Analysis for large scale omics data integration, biomarker discovery, drug repositioning and screening for new therapeutic targets for Alzheimer's Disease

Patel, Hamel

Awarding institution:
King's College London

Analysis for large scale omics data integration, biomarker discovery, drug repositioning and screening for new therapeutic targets for Alzheimer's Disease

Hamel Patel

Social, Genetics and Developmental Psychiatry Centre

Institute of Psychiatry, Psychology and Neuroscience

King's College London

A thesis submitted to King's College London

for the degree of Doctor of Philosophy

November 2019

Abstract

The increase in life expectancy has profoundly increased the ageing population, which, unfortunately, is also accompanied by a rise in age-related disorders, including Alzheimer's Disease (AD). The most common form of dementia is AD, which was first described over a century ago, however, to date, there still exists a lack of understanding the molecular changes specific to the disease, a clinically established robust blood-based biomarker for accurate disease diagnosis and a lack of treatments.

This thesis begins by investigating microarray gene expression profiling from Asymptomatic AD (AsymAD) human brains, who were clinically free from dementia; however, upon autopsy, they were observed to consist of hallmark AD pathology. A significant increase of transcriptomic activity in the frontal cortex (FC) brain region of AsymAD subjects was detected, suggesting fundamental changes in AD may initially begin within the FC brain region prior to symptoms of AD. In addition, overactivation of the brain "glutamate-glutamine cycle" and disruption to the brain energy pathways in both AsymAD and AD subjects were identified, suggesting these may be the earliest biological pathways disrupted in the disease, providing potential targets for early disease intervention.

Secondly, existing and novel microarray gene expression studies of human AD brains were integrated into the largest known AD meta-analysis to date and is the first to incorporate numerous non-AD neurological disorders to identify AD-specific molecular changes. Seven genes were observed to be specifically and consistently perturbed across AD brains, with SPCS1 gene expression pattern found to replicate in RNA-Seq data. The cerebellum brain region is often regarded to be free from hallmark AD pathology and was incorporated into the analysis as a secondary control to identify an additional nineteen genes that may be involved with hallmark AD pathology. Furthermore, biological processes often reported as disrupted in AD were observed to be tissue-specific, and viral components were found to be specifically enriched across AD brains.

Thirdly, an automated transcriptomic based drug repositioning pipeline was developed to query the reprocessed Connectivity Map to identify candidate compounds for disease intervention. Drug repositioning the AsymAD gene expression profile identified several candidate compounds that are already FDA approved for the treatment of AD and cognitive impairment, suggesting these compounds may be effective in the early stage of the disease. Drug repositioning the AD gene expression profile identified an anti-biotic compound for disease intervention.

Finally, a machine learning approach was used to identify a blood-based 28 gene expression profile, which is enriched for "herpes simplex infection", and can distinguish AD from Parkinson's Disease,

Multiple Sclerosis, Amyotrophic Lateral Sclerosis, Bipolar Disorder, Schizophrenia, Coronary Artery Disease, Rheumatoid Arthritis, Chronic Obstructive Pulmonary Disease, and cognitively healthy subjects with 66.3% PPV and 90.6% NPV.

Overall, the work undertaken in this thesis provides new insight into the molecular changes occurring in both the asymptomatic and symptomatic phase of the disease, demonstrates a framework for a possible blood-based transcriptomic diagnosis test, provides new potential therapeutic targets, identifies candidate compounds that require further investigation for disease intervention and provides new avenues for future AD-related research.

Acknowledgements

I would like to thank my supervisor's Professor Richard Dobson and Dr Stephen Newhouse, for providing me with this opportunity and for their continuous support and advice throughout this work. I sincerely appreciate the time you both have provided me.

I cannot begin to express my gratitude towards my family for the endless love, support and encouragement they have provided me throughout this journey. A special thank you to my wonderful wife Dolly for the emotional support and the countless sacrifices made to help me get to this point, and my two mischievous children Milly and Rubal, who have provided the much-needed joyful distractions and laughter throughout this work. Thank you all!

Statement of Authorship

All work included in this thesis, both analysis and written, was performed by Hamel Patel with the following exceptions:

1. The R Shiny app in Chapter 2 was developed and deployed by Hamel Patel. The web interface was later refined by Steven J Newhouse.
2. Chapter3: “A Meta-analysis of Alzheimer’s Disease Brain Transcriptomic Data” is a publication which was circulated for co-author review and was subject to peer review prior to publication.
3. The samples in all datasets used in this thesis were collected and expression profiled by others. In all cases, this is made clear in the methodology sections of each chapter.

List of Abbreviations

Abbreviation	Details
AD	Alzheimer's Disease
ADAS-Cog	AD Assessment Scale-Cognitive Subscale
ADIPOR1	Adiponectin Receptor 1
ALS	Amyotrophic Lateral Sclerosis
AMP-AD	Accelerating Medicines Partnership-Alzheimer's Disease
APOE	Apolipoprotein E
APP	Amyloid Precursor Protein
AQCg	Accuracy Quality Control Gene Index
AQCp	Accuracy Quality Control Pathway Index
AsymAD	Asymptomatic AD
ATC	Anatomic Therapeutic Chemical
AUC	Area Under Curve
AW	Adaptively Weighted
AW.OC	Adaptively Weighted with One-sided Correction
Aβ	Amyloid-Beta
BACE	β-site Amyloid precursor protein cleaving enzyme 1
BBB	Blood-Brain Barrier
BD	Bipolar Disorder
BNIP3L	B-cell 2 Interacting Protein 3 Like
BRC-MH	Biomedical Research Centre for Mental Health
CB	Cerebellum
CD	Coronary Artery Disease
CDR	Clinical Dementia Rating Scale
CJD	Creutzfeldt-Jakob Disease
CMap	Connectivity Map
CNS	Central Nervous System
COPD	Chronic Obstructive Pulmonary Disease
CQCg	Consistency Quality Control Gene Index
CQCp	Consistency Quality Control Pathway Index
CSF	Cerebral Spinal Fluid
CUI	Clinical Utility Index

CV	Cross-validation
CVD	Cardiovascular Disease
DE	Differentially Expressed
DEG's	Differentially Expressed Genes
DLB	Dementia with Lewy Bodies
EC	Entorhinal Cortex
EOAD	Early-onset AD
EQC	External Quality Control Index
FC	Frontal Cortex
FDA	Food and Drug Administration
FDG	Fluoro-2-Deoxy-D-glucose
FDR	False Discovery Rate
fMRI	Functional Magnetic Resonance Imaging
FTD	Frontotemporal Dementia
FTLD	Frontotemporal Lobar Degeneration
GABA	gamma-aminobutyric acid
GEO	Gene Expression Omnibus
GO	Gene Ontology
GPUs	Graphics Processing Units
GSEA	Gene Set Enrichment Analysis
GWAS	Genome-Wide Association Studies
HD	Huntington's Disease
HPC	High-Performance Computing
HSV1	Herpes Simplex Virus Type 1
IoPPN	Institute of Psychiatry, Psychology and Neuroscience
IQC	Internal Quality Control Index
IQR	inter-quartile range
KDM5D	Lysine Demethylase 5D
KS	Kolmogorov–Smirnov
Limma	Linear Models for Microarray Data
LINCS	Library of Integrated Network-based Cellular Signatures
LOAD	Late-onset AD
MAP	Microtubule-Associated Protein
massiR	MicroArray Sample Sex Identifier

MBCB	Model-Based Background Correction
MCF7	Human breast epithelial adenocarcinoma cell line
MCI	Mild Cognitive Impairment
MDD	Major Depressive Disorder
MKRN1	Makorin Ring Finger Protein 1
ML	Machine Learning
MMSE	Mini-Mental State Examination
MOSPD3	Motile Sperm Domain Containing 3
MRC	Medical Research Council
MRI	Magnetic Resonance Imaging
MS	Multiple Sclerosis
MVAD	Mixed Vascular-Alzheimer Dementia
NFT	Neurofibrillary Tangles
NMDA	N-Methyl-D-Aspartate
NOTCH2NL	Notch 2 N-Terminal
NPC2	NPC Intracellular Cholesterol Transporter 2
NPV	Negative Predictive Power
OGT	O-Linked N-Acetyl Glucosamine Transferase
PC's	Principal Components
PCA	Principal Component Analysis
PD	Parkinson's Disease
PET	Positron Emission Tomography
PL	Parietal Lobe
PM	Post-Mortem
PPI	Protein-Protein Interactions
PPV	Positive Predictive Power
PRKY	Protein Kinase Y-Linked
ProbCMap	Probabilistic Connectivity Map
PRS	Polygenic Risk Scores
PSEN1	Presenilin 1
PSEN2	Presenilin 2
QC	Quality Control
QCg	Quality Control Genes
RA	Rheumatoid Arthritis

RNA-Seq	RNA-sequencing
RPS4Y1	Ribosomal Protein S4 Y-Linked 1
RRHO	Rank-Rank Hypergeometric Overlap
RSN	Robust Spline Normalisation
RT-PCR	Polymerase Chain Reaction
SCZ	Schizophrenia
SE	Standard Error
SLaM	South London and Maudsley NHS Foundation Trust
sMRI	structural Magnetic Resonance Imagines
SOD1	Superoxide Dismutase 1
SPCS1	Signal Peptidase Complex Subunit 1
SPIA	Signalling Pathway Impact Analysis
SPIED	Searchable Platform-Independent Expression Database
ssCMap	Statistically significant Connectivity Map
SUMO2	Small Ubiquitin-like Modifier 2
SVA	Surrogate Variable Analysis
TC	Temporal Cortex
TRIL	TLR4 Interactor with Leucine-Rich Repeats
TTD	Therapeutic Targets Database
UBC	Ubiquitin-C
UPS	Ubiquitin-Proteasome System
USP9Y	Ubiquitin Specific Peptidase 9 Y-Linked
UTY	Ubiquitously Transcribed Tetratricopeptide Repeat Containing, Y-Linked
VaD	Vascular Dementia
WGCNA	Weighted Gene Correlation Network Analysis
XGBoost	Extreme Gradient Boosting
XIST	X Inactive Specific Transcript

Table of Contents

List of Figures

Publications arising from this thesis

Hamel Patel, Richard J.B Dobson, Stephen J. Newhouse (2019) "A Meta-Analysis of Alzheimer's Disease Brain Transcriptomic Data", Journal of Alzheimer's Disease 68(4):1635-1656.

Hamel Patel, Angela K. Hodges, Charles Curtis, Sang Hyuck Lee, Claire Troakes, Richard J.B Dobson, Stephen J. Newhouse "Transcriptomic analysis of probable asymptomatic and symptomatic Alzheimer brains", Brain, Behaviour, and Immunity 80:644-656.

Additional Publications arising in parallel to this thesis

Deliah Zaganeh, Eva Krapohl, Helena Alexandra Gaspar, Charles Curtis, Sang Hyuck Lee, **Hamel Patel**, Stephen J. Newhouse, H. M. Wu, Michael A Simpson, Martha Putallaz, David Lubinski, Robert Plomin, and Gerome. Breen (2018) "A Genome-Wide Association Study for Extremely High Intelligence." Molecular Psychiatry 23(5):1226–32.

Marcos Leite Santoro, Vanessa Ota, Simone de Jong, Cristiano Noto, Leticia M. Spindola, Fernanda Talarico, Eduardo Gouvea, Sang Hyuck Lee, Patricia Moretti, Charles Curtis, **Hamel Patel**, Stephen Newhouse, Carolina Muniz Carvalho, Ary Gadelha, Quirino Cordeiro, Rodrigo Affonseca Bressan, Sintia Iole Belangero, and Gerome Breen (2018) "Polygenic Risk Score Analyses of Symptoms and Treatment Response in an Antipsychotic-Naive First Episode of Psychosis Cohort." Translational Psychiatry 8(1):174.

Eva Krapohl, **Hamel Patel**, Stephen J. Newhouse, Charles Curtis, Sophie Von Stumm, Philip Dale, Deliah Zabaneh, Gerome Breen, Paul. F. O'Reilly, and Robert Plomin (2018) "Multi-Polygenic Score Approach to Trait Prediction." Molecular Psychiatry 23(5):1368–74.

Kate Gardner, Tony Fulford, Nicholas Silver, Helen Rooks, Nikolaos Angelis, Marlene Allman, Siana Nkya, Julie Makani, Jo Howard, Rachel Kesse-Adu, David C. Rees, Sara Stuart-Smith, Tullie Yeghen, Moji Awogbade, Raphael Z. Sangeda, Josephine Mgaya, **Hamel Patel**, Stephen Newhouse, Stephan Menzel, and Swee Lay Thein. n.d (2018) "G(HbF): A Genetic Model of Fetal Hemoglobin in Sickle Cell Disease." Blood advances 2(3): 235-239.

Chiara Fabbri, Katherine Tansey, Roy Perlis, Joanna Hauser, Neven Henigsberg, Wolfgang Maier, Ole Mors, Anna Placentino, Marcella Rietschel, Daniel Souery, Gerome Breen, Charles Curtis, Lee Sang-Hyuk, Stephan Newhouse, **Hamel Patel**, Michel Guipponi, Nader Perroud, Guido Bondolfi, Michael O'Donovan, Glyn Lewis, Joanna Biernacka, Richard Weinshilboum, Anne Farmer, Katherine Aitchison, Ian Craig, Peter McGuffin, Rudolf Uher, and Catheryn M. Lewis (2018) "New Insights into the

Pharmacogenomics of Antidepressant Response from the GENDEP and STAR*D Studies: Rare Variant Analysis and High-Density Imputation." The Pharmacogenomics Journal 18(3):413–21.

Nicola Voyle, Hamel Patel, Amos Folarin, Stephen Newhouse, Caroline Johnston, Pieter Jelle Visser, Richard JB. Dobson and Steven J. Kiddle for the Alzheimer's Disease Neuroimaging Initiative (2017) Genetic risk and plasma tau as markers of amyloid-β and tau burden in cerebrospinal fluid. Journal of Alzheimer's Disease. 55(4):1417-1427

Evangelos Vassos, Marta Di Forti, Jonathan Coleman, Conrad Iyegbe, Diana Prata, Jack Euesden, Paul O'Reilly, Charles Curtis, Anna Kolliakou, Hamel Patel, Stephen Newhouse, Matthew Traylor, Olesya Ajnakina, Valeria Mondelli, Tiago Reis Marques, Poonam Gardner-Sood, Katherine J. Aitchison, John Powell, Zerrin Atakan, Kathryn E. Greenwood, Shubulade Smith, Khalida Ismail, Carmine Pariante, Fiona Gaughran, Paola Dazzan, Hugh S. Markus, Anthony S. David, Cathryn M. Lewis, Robin M. Murray, and Gerome Breen. (2017) "An Examination of Polygenic Score Risk Prediction in Individuals With First-Episode Psychosis." Biological Psychiatry 81(6):470–77.

Matthew Traylor, Charles Curtis, Hamel Patel, Gerome Breen, Sang Hyuck Lee, Xiaohui Xu, Stephen Newhouse, Richard Dobson, Sophia Steer, Andrew P. Cope, Hugh S. Markus, Cathryn M. Lewis, and Ian C. Scott (2017) "Genetic and Environmental Risk Factors for Rheumatoid Arthritis in a UK African Ancestry Population: The GENRA Case-control Study." Rheumatology 56(8):1282–92.

Matthew Traylor, Loes Rutten-Jacobs, Charles Curtis, Hamel Patel, Gerome Breen, Stephen Newhouse, Cathryn M. Lewis, and Hugh S. Markus (2017) "Genetics of Stroke in a UK African Ancestry Case-Control Study." Neurology Genetics 3(2):e142.

Rebecca Harrison, Fiona Gaughran, Robin M. Murray, Sang Hyuck Lee, Jose Paya Cano, David Dempster, Charles J. Curtis, Danai Dima, Hamel Patel, Simone de Jong, and Gerome Breen (2017) "Development of Multivariable Models to Predict Change in Body Mass Index within a Clinical Trial Population of Psychotic Individuals." Scientific Reports 7(1):14738.

García-González, Judit, Katherine E. Tansey, Joanna Hauser, Neven Henigsberg, Wolfgang Maier, Ole Mors, Anna Placentino, Marcella Rietschel, Daniel Souery, Tina Žagar, Piotr M. Czerski, Borut Jerman, Henriette N. Buttenschøn, Thomas G. Schulze, Astrid Zobel, Anne Farmer, Katherine J. Aitchison, Ian Craig, Peter McGuffin, Michel Giupponi, Nader Perroud, Guido Bondolfi, David Evans, Michael O'Donovan, Tim J. Peters, Jens R. Wendland, Glyn Lewis, Shitij Kapur, Roy Perlis, Volker Arolt, Katharina Domschke, Gerome Breen, Charles Curtis, Lee Sang-Hyuk, Carol Kan, Stephen Newhouse, Hamel Patel, Bernhard T. Baune, Rudolf Uher, Cathryn M. Lewis, and Chiara Fabbri (2017) "Pharmacogenetics of

Antidepressant Response: A Polygenic Approach." Progress in Neuro-Psychopharmacology and Biological Psychiatry 75:128–34.

Simone de Jong,, Stephen J. Newhouse, Hamel Patel, Sanghyuck Lee, David Dempster, Charles Curtis, Jose Paya-Cano, Declan Murphy, C. Ellie Wilson, Jamie Horder, M. Andreina Mendez, Philip Asherson, Margarita Rivera, Helen Costello, Stefanos Maltezos, Susannah Whitwell, Mark Pitts, Charlotte Tye, Karen L. Ashwood, Patrick Bolton, Sarah Curran, Peter McGuffin, Richard Dobson, and Gerome Breen (2016) "Immune Signatures and Disorder-Specific Patterns in a Cross-Disorder Gene Expression Analysis." British Journal of Psychiatry 209(03):202–8.

Jonathan Coleman, Jack Euesden, Hamel Patel, Amos A. Folarin, Stephen Newhouse, and Gerome Breen (2016) "Quality Control, Imputation and Analysis of Genome-Wide Genotyping Data from the Illumina HumanCoreExome Microarray." Briefings in Functional Genomics 15(4):298–304.

Chapter 1: Introduction

1.1 Alzheimer's disease

In 1906, a clinical psychiatrist and neuroanatomist named Alois Alzheimer first described a 50-year-old patient with memory issues, paranoia, disorientation, and deteriorating psychological changes until her death. Post mortem examination identified dramatic changes to the brain, including shrinkage and abnormal altercations. Alois identified a further three patients with similar clinical and histopathological symptoms, and by 1909, the illness was referred to as Alzheimer's disease (AD) (Hippius and Neundörfer, 2003).

AD falls under the umbrella term dementia, which is characterised by loss of cognitive function, behavioural changes and social functioning. An estimated 50 million people worldwide are living with dementia (Patterson, 2018), of which AD is estimated to account for 60-80% (Gaugler *et al.*, 2016), making it the most common cause of dementia. In the UK alone, an estimated 200,000 new cases of dementia are diagnosed every year (Matthews *et al.*, 2016), which has a significant economic impact, costing an approximate £26 billion to care for patients and has been forecasted to double by 2050 (Lewis *et al.*, 2014).

1.1.1 Disease progression

Ageing is an ongoing natural process of living, where structural, biochemical, physiological and functional changes lead to increased vulnerability and frailty. As everyone is expected to experience cognitive decline during ageing while exhibiting different symptoms, the transition from normal ageing to AD is very subtle and is modelled in Figure 1.1 (Sperling *et al.*, 2011). Cognitive decline can be measured by several different tests and generally involves a series of questions to assess an individual's short and long term memory, concentration, attention, language and orientation (Hunt *et al.*, 2017).

The rate of decline varies between AD subjects and has been suggested to take a few years to a couple of decades for dementia to manifest (Thalhauser and Komarova, 2012). As the disease progresses, the brain progressively becomes damaged, with the severity of symptoms a typical reflection of the degree of neuronal loss. Individuals with the disease may exhibit multiple symptoms that change over several years. In addition, everyone may not experience the condition in the same manner, making symptoms challenging to distinguish from healthy ageing and other disorders such as depression (Gaugler *et al.*, 2016). Nevertheless, AD can be generalised into three distinct stages based on symptoms; the preclinical phase, mild cognitive impairment (MCI) and dementia due to AD.

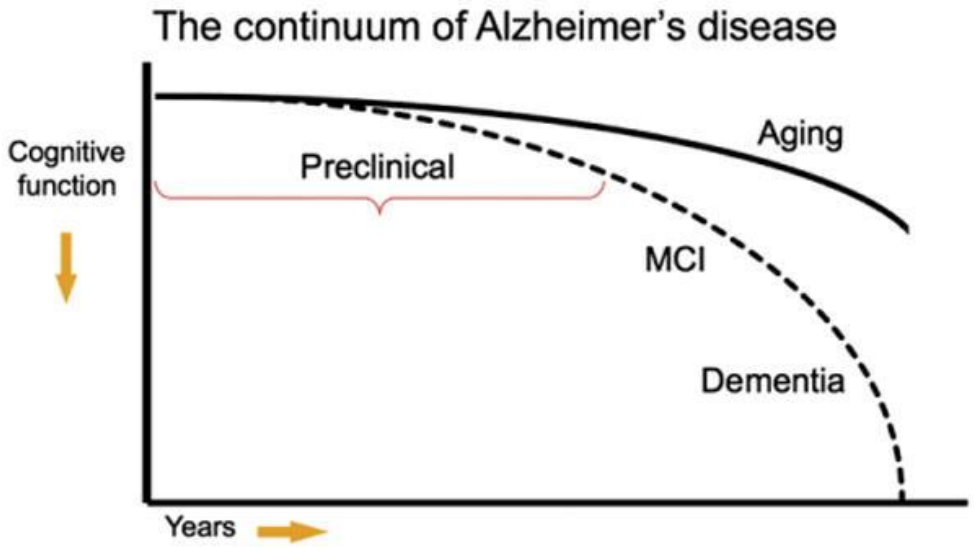


Figure 1.1: Model of clinical trajectory of AD with three distinct stages based on symptoms; 1) preclinical stage, 2) MCI and 3) Dementia due to AD. Reproduced from (Sperling et al., 2011).

1.1.2 Genetic causes

AD can be generalised as either early-onset AD (EOAD), where the disease manifests before the age of 65 and is inherited in a Mendelian dominant fashion, or late-onset AD (LOAD), where the disease manifests after the age of 65 and is driven by a combination of genetic and environmental factors. Approximately 95% of all AD cases are LOAD, while 5% are EOAD. No single genetic mutation has been identified to account for all AD cases. However, mutations in either the amyloid precursor protein (APP), presenilin 1 (PSEN1) or presenilin 2 (PSEN2) gene will result in EOAD.

1.1.3 Risk factors

A number of genetic and environmental factors have been found to increase the risk of AD. These factors increase the risk of developing AD but do not necessarily imply an individual with these risk factors will be guaranteed to develop the disease.

1.1.3.1 Non-genetic risk factors

The most significant risk factors for LOAD are age and gender. The prevalence of AD significantly increases with age and females are at greater risk. A typical 65-year old female is estimated to be at 12% risk, while a male of the same age has a risk of 6.3% (Podcasy and Epperson, 2016). The difference in disease prevalence in genders has been suggested to be attributed to women's longevity in combination with young female mitochondria being protected against amyloid-beta (Aβ) toxicity, generating less reactive oxygen species, and releasing less apoptogenic signals than from males (Viña, Viña and Lloret, 2010).

Approximately 25% of all AD patients have a family history of AD. Individuals whose parents or siblings have AD have an increased likelihood to develop the disease and are at even more risk if they have more than one first-degree relative with AD (Bird, 2018).

The risk factors associated with cardiovascular disease (CVD), such as smoking, obesity, diabetes and hypertension, have also been shown to increase the risk of AD. Although the direct connection between the two diseases remains unclear, the general hypothesis suggests a healthy brain's demand for energy supply is hampered by CVD, resulting in decreased nutrient and oxygen-rich blood reaching the brain for normal functionality, causing nerve cell death.

Individuals who have more years in education, a mentally stimulating occupation or engage in social activities have been suggested to build "cognitive reserves" and reduce the risk of AD (Wang, Xu and Pei, 2012; Iacono *et al.*, 2015; Grzywacz, Segel-Karpas and Lachman, 2016). These "cognitive reserves" are the brain's ability to create flexible cognitive neuronal networks, which have been suggested to optimise normal cognition and compensate for brain changes such as accumulation of Aβ and tau, enabling an individual to continue to carry out cognitive tasks (Stern, 2002).

Traumatic brain injury (TBI) is a blow, jolt or penetration of the skull caused by a foreign object, which disrupts normal brain function and can result in a loss of consciousness and post-traumatic amnesia. Studies have generated a growing body of evidence to suggest that TBI's and repeated head injuries (e.g. boxers, and football players) increase the risk of MCI, dementia and neurodegenerative diseases including AD (Guskiewicz *et al.*, 2005; Davis *et al.*, 2015).

1.1.3.2 Genetic risk factors

Multiple susceptibility genes have been shown to increase the risk of LOAD, with the APOE gene, located on chromosome 19, being the most significant. The APOE gene has three major allelic variants (e2, e3, and e4), which encode for different isoforms of a cholesterol transporting protein. The e3 variant is the most common, followed by the e4 and then the e2 variant (Mahley and Rall, 2000). An individual inherits one APOE gene from each parent and therefore, can have any one of the six different allele combinations (e2/e2, e2/e3, e2/e4, e3/e3, e3/e4, e4/e4). Individuals who inherit one copy of the e4 have a 3-fold increase in developing AD, while inheriting two copies of the e4 form increases the risk to 15-fold when compared to two copies of the e3 form. Approximately 25% of the general population and up to 65% of AD patients have the e4 variant (Bird, 2018). In contrast, inheriting the e2 variant may decrease the risk of developing AD when compared with the e3 variant (Iacono *et al.*, 2015).

An additional nineteen variants have been identified to increase the risk of AD, but collectively only account <2% of all cases. These genes are ABCA7, AKAP9, BIN1, CASS4, CD2AP, CD33, CLU, EPHA1,

FERMT2, HLA-DRB5/DRB1, INPP5D, MEF2C, MS4A6A/MS4A4E, PICALM, PLD3, PTK2B, SORL1, TREM2 and UNC5C. Many of these genes are involved in various brain development, cytoskeleton organisation, and immune functions (Bird, 2018).

1.1.4 Current understanding of the disease pathology

Although the exact cause of the disease remains a mystery, many hypotheses exist, with amyloid and tau hypotheses arguably being the most common. Two proteins in the brain are associated with these hypotheses. The first is Amyloid Beta (Aβ), which accumulates between brain neurons to form insoluble structures known as plaques, and the second is tau, which collectively forms neurofibrillary tangles inside neurons. Plaques and tangles exist in all older adults; however, there is an abnormal accumulation in individuals with AD, which has been suggested to be a result of an imbalance in the production and removal of these proteins. Evidence has even suggested the abnormal accumulation of these proteins occur in AD brains 20 years prior to the onset of clinical symptoms (Reiman *et al.*, 2012).

1.1.4.1 The Amyloid hypothesis

The Amyloid hypothesis assumes the abnormal accumulation of brain Aβ forms plaques, which trigger neurodegeneration and ultimately is the cause of AD (Makin, 2018). Aβ is a peptide consisting of 40 amino acids, which is derived from sequential cleavage of the Amyloid Precursor Protein (APP). The APP is a single-pass transmembrane protein that is highly expressed in brain neurons. The exact function of APP is unknown; however, studies suggest APP modulates cell growth, promotes neuronal survival, neurite outgrowth and is involved in general cell health (O'Brien and Wong, 2011).

In normal processing, the APP is sequentially cleaved into several peptides by enzyme activities. APP is first cleaved within the lumenal domain by β- secretase (BACE1) or α-secretase (complex consisting of presenilin), resulting in a membrane-tethered β- or α- C-terminal fragments respectively. These fragments are further cleaved by γ-secretase to release $A\beta_{40}$ or $A\beta_{42}$ residues outside the cell.

As illustrated in Figure 1.2, Aβ can be cleared by astrocytes, or they aggregate to form Aβ-plaques, which can be cleared by macrophage degradation, phagocytic clearance by microglia or by endoproteases from astrocytes. However, some oligomers dissociate from the Aβ plaques to form insoluble forms of Aβ plaques which cannot be cleared. These plaques primarily consist of the $A\beta_{42}$ isoform, and it interferes with neighbouring synapses to induce tau aggregation in neurons by unknown mechanisms, which causes neuronal damage (Masters *et al.*, 2015).

Plaque burden has been found to be weakly correlated to the disease severity, and since plaques are naturally found in the brains of the elderly with normal cognition, the amyloid hypothesis has been suggested not to be the primary cause of AD (Makin, 2018). Although, there is counter-argument that

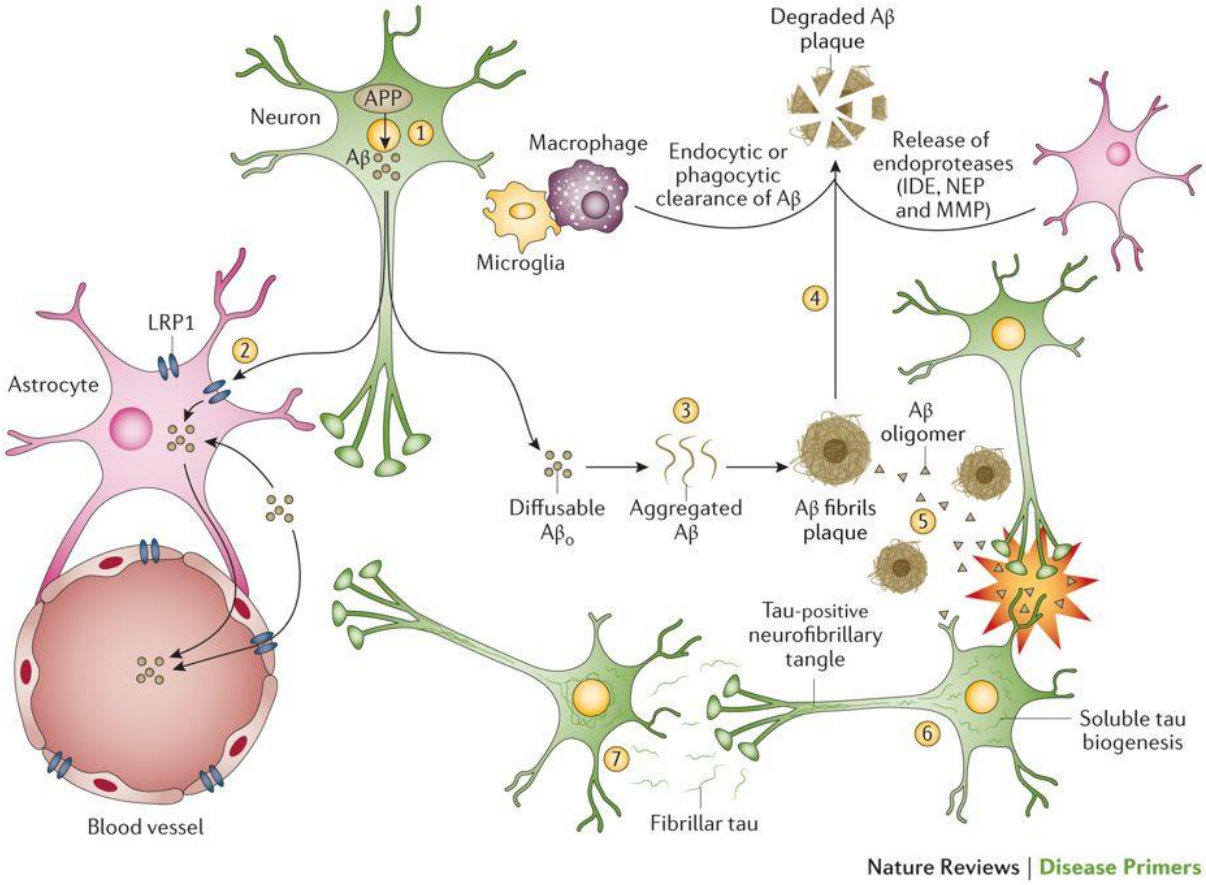


Figure 1.2: Pathway leading to the accumulation of plaques and tangles form the basis of AB theory. Aß is cleaved from an amyloid precursor protein (APP; step 1) and is released into the extracellular milieu — by a process that is unclear — as diffusible oligomers (Aßo). Aßo can be cleared by mechanisms that involve APOE or can be taken up by astrocytes via low-density lipoprotein receptor-related protein 1 (LRP1; step 2). Aßo can also aggregate in the intercellular space to form fibrillary constructs, which in turn assemble into plaques (step 3). Aß plaques can be cleared from the brain via degradation by endocytic or phagocytic clearance (in macrophages and microglia), or by endoproteases from astrocytes (such as an insulin-degrading enzyme (IDE), neprilysin (NEP) and matrix metalloproteinase (MMP); step 4). However, some conformational oligomers that dissociate from Aß fibrils and plaques may not be cleared and are toxic to adjacent synapses (step 5) and induce tau aggregation by as yet unknown mechanisms. Tau damage occurs in neurons and is mediated by the development of tau-positive neurofibrillary tangles (which extend into the dendrites; step 6). Fibrillar tau can be released and taken up by healthy neurons, triggering tau damage in the up taking cell (step 7). In addition, Aß oligomers might drive α-synuclein aggregation in the plaques. Reproduced from (Masters *et al.*, 2015).

these subjects may be in the presymptomatic stage of AD, where the accumulation of plaques is not yet sufficient to cause enough neuronal disruption for the appearance of clinical symptoms (Makin, 2018).

1.1.4.2 The Tau hypothesis

Cognitive symptoms of AD are more correlated with the severity and location of tau pathology compared to Aβ plaques. Furthermore, human autopsy examinations have observed neurofibrillary tangles (NFT) formation without Aβ plaques, suggesting these two proteins may arise independently in

AD or alternatively tangles may occur prior to amyloid plaques (Makin, 2018), fuelling NFT as the main cause of AD and giving rise to the tau hypothesis.

Microtubule-associated protein (MAP) bind to the tubulin part of microtubules and regulate their organisation and turnover. There are a number of different MAPs which either stabilise microtubules to promote polymerisation or inhibit their depolymerisation. Binding of other MAPs can cause instability to the microtubule structure, preventing their assembly (Cassimeris, 2009). Tau is one of the MAPs that regulate the stability of tubulin assembly of normal mature neurons through phosphorylation. Six isoforms of the tau gene on chromosome 17 are found in the human brain, generated by alternative splicing of the pre-mRNA (Iqbal *et al.*, 2010). Normal brain tau contains 2-3 moles of phosphate per mole of the protein for optimal functionality; however, in AD, tau is 3-4-fold more phosphorylated (Köpke *et al.*, 1993). Although the exact process is unknown, these hyperphosphorylated tau proteins accumulate as straight filaments or unique twisted fibrils to form an insoluble structure in neuronal bodies and are referred to as NFT (Kametani and Hasegawa, 2018). These NFTs have been suggested to cause synaptic dysfunction and neuronal death, leading to neurodegeneration.

Brain tau pathology is measured using BRAAK staging, where BRAAK stages I-II represent early appearance in the entorhinal region, BRAAK stages III-IV refers to the compromise of the limbic region, and BRAAK stages V-VI are assigned when the neocortical areas are affected (Braak and Braak, 1991). In contrast to Aβ, several studies have identified a strong correlation of BRAAK with cognitive decline in AD, further implying the tau hypothesis as a causative role (Kametani and Hasegawa, 2018).

1.1.4.3 Alternative AD pathology hypothesis

The Amyloid and Tau hypotheses have been the mainstream causative concepts for AD; however, therapies targeting the accumulation of Aβ plaques and NFT have clinically failed, leading to suggestions of alternative causative theories (Kametani and Hasegawa, 2018; Makin, 2018). In addition to the accumulation of Aβ plaques and NFT, oxidative stress and inflammation have been reported in AD brains. A study suggested abnormalities in both the mitotic signalling and oxidative stress signalling pathways may propagate disease pathogenesis by releasing growth factors that activate the Ras-ERK pathway in susceptible neurons, resulting in the re-entry of the cell cycle and tau phosphorylation (Zhu *et al.*, 2004).

Brain mitochondrial dysfunction has been reported in AD on numerous occasions and combined with the strong maternal genetic contribution in AD, give rise to the mitochondrial cascade hypothesis (Swerdlow and Khan, 2004). The theory suggests that mitochondrial dysfunction contributes to

elevated ROS by generating Aβ, resulting in neuronal progenitors unsuccessfully entering cell-cycle and causing phosphorylation of tau and NFT formation.

Several studies have also suggested that genetic mutations observed in AD increases an individual's susceptibility to infections and viruses (Itzhaki *et al.*, 1997; Porcellini *et al.*, 2010), leading to the viral hypothesis. The concept suggests viruses travel to the brain by middle age due to the age-related weakening of the immune system, where it remains in a latent form until the virus reactivates by an unknown trigger. Herpes Simplex Virus Type 1 (HSV1) is one of many viruses observed in the brains of AD patients (De Chiara *et al.*, 2012; Itzhaki, 2018; Readhead *et al.*, 2018). In addition, Aβ is an innate immune protein that protects against viral infections and has been shown to bind glycoproteins of the HSV, accelerating Aβ deposition and resulting in protective entrapment of the virus (Eimer *et al.*, 2018). The viral hypothesis suggests Aβ may play a protective role in the brain, and brain viral infections may directly promote Aβ amyloidosis, however, like the Amyloid hypothesis, the viral hypothesis cannot currently explain why NFT may be observed prior to Aβ plaques.

1.1.5 The neuropathological spread of disease.

The brain is a complex network consisting of over 100 billion neurons, each communicating through end terminals known as synapses. Over 100 trillion synapses exist in the brain, which allows information to pass in the form of small chemicals to create the cellular basis of memories, thoughts, sensation, emotions and movement (Gaugler *et al.*, 2016). As illustrated in Figure 1.3, the brain can be compartmentalised into three regions; the cerebrum, cerebellum and brainstem.

The cerebrum, which is also referred to as the cerebrum cortex, is the largest part of the brain and is composed of the right and left hemispheres, each controlling the opposite side of the body. Each hemisphere does not share functions, and in general, the left hemisphere is dominant for hand use and language while the right hemisphere controls creativity, spatial ability, artistic and musical skills. The cerebral hemispheres can be further compartmentalised into the temporal lobe, frontal lobe, parietal lobe and occipital lobe, with each being specialised in functionality.

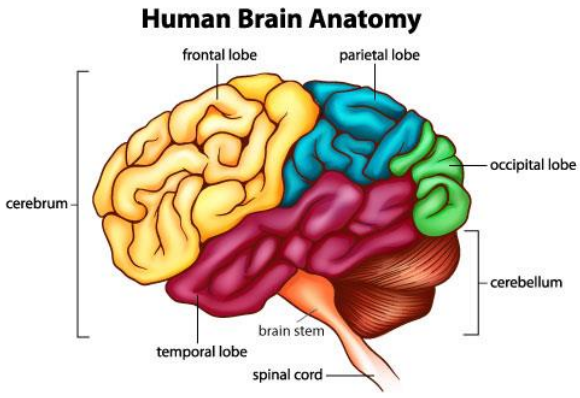


Figure 1.3: Basic anatomy of the human brain. Reproduced from http://idahoptv.org/sciencetrek/topics/brain/facts.cfm 01/11/2019

Nerve cell death, tissue loss and atrophy occur throughout the brain as AD progresses, leading to the manifestation of clinical symptoms associated with loss of normal brain function. Although the exact

cause of pathology remains unknown, the general spread through the human brain occurs in a predictable pattern. In the initial phase of the disease, Aβ plaques and NFT's accumulates in the cerebrum, specifically the hippocampus and entorhinal cortex which are housed in the temporal lobe and both primarily associated with learning and memory (Serrano-Pozo *et al.*, 2011). Next, the frontal lobe is affected, a region associated with reasoning, motor skills, higher levels of cognition and expressive language, followed by the Parietal lobe, a region involved in processing sensory information such as pressure, touch and pain. In later stage AD, the occipital lobe can become affected, a region involved with the interpretation of visual information from the eyes.

Eventually, the brain stem, which is involved in body movement and vital to autonomic functions such as blood pressure and breathing are affected, followed by the cerebellum. The cerebellum only accounts for 10% of the brain but contains over 50% of the brains total neurones and is often referred to as the "little brain". The cerebellum is involved in motor movement and control, specifically posture, balance and coordination of voluntary movements. The cerebellum is generally considered to be partially spared from the disease as plaques are only occasionally seen, and tangles are generally not reported (Convit et al. 2000; (Jacobs *et al.*, 2017).

1.1.6 Symptoms and clinical diagnosis

To date, there is no definitive diagnostic test for AD. The current clinical tests only suggest AD as a likely suspect, with confirmation only possible at autopsy (Bird, 2018). Establishing a clinical diagnosis of AD requires a time-consuming comprehensive combination of physical, mental and neuropsychological examinations.

The National Institute on Aging and the Alzheimer's Association have guidelines for diagnosing AD at three distinct stages; preclinical AD, MCI due to AD and Dementia due to AD. Diagnosis of dementia is relatively simple and does not require any specialised brain scans, however, identifying the underlying cause is exceptionally complex as dementia can be caused by a number of different diseases and non-diseased factors, such as AD, Vascular Dementia (VaD), Dementia with Lewy bodies (DLB), Parkinson's Disease (PD), Frontotemporal Dementia (FTD), depression, drug intoxication, thyroid disease, vitamin deficiency, CNS infections and normal-pressure hydrocephalus (Gaugler *et al.*, 2016).

1.1.6.1 Preclinical AD

Brain changes associated with AD have been suggested to occur up to 20 years prior to the appearance of clinical symptoms (Reiman *et al.*, 2012). An estimated 20-30% of the ageing population with intact cognition have AD neuropathology, and it has been postulated these individuals are more likely to develop AD (Yvette *et al.*, 2010). These individuals are often referred to as asymptomatic AD (AsymAD) or preclinical AD (Driscoll and Troncoso, 2011). Preclinical AD precedes MCI but does not necessarily

mean individuals will develop AD or even MCI (Sperling *et al.*, 2011), nevertheless, identifying these patients can lead to early disease intervention such as the implementation of coordinated care plans, better management of symptoms and patient safety (Dubois *et al.*, 2015). Preclinical AD can be diagnosed primarily for research purposes when subtle changes in cognitive decline are apparent but do not meet the criteria of MCI, carry one or more apolipoprotein E (APOE) e4 variant or are carriers of autosomal dominant mutations, which will almost certainly develop AD (Sperling *et al.*, 2011).

1.1.6.2 MCI

MCI is generally an intermediate stage between cognitive decline due to normal ageing and cognitive decline due to dementia. There is no single cause or outcome for MCI; however, the risk of developing dementia, including AD, is increased. MCI is a heterogeneous condition with 16 different classifications used to define it (Arevalo-Rodriguez *et al.*, 2015). In general, MCI is a state where individuals may exhibit mild memory complaints but have intact cognitive functions with little adverse effects on their daily life. An estimated 15-20% of 65-year olds or older have MCI (Roberts and Knopman, 2013), who over time, may experience a gradual cognitive decline and change in personality and behaviour. An estimated 33% of MCI have been reported to develop dementia within five years, while up to 10% do not progress to dementia in 10 years (Mitchell and Shiri-Feshki, 2009).

Mild cognitive impairment is difficult to distinguish from healthy ageing and dementia due to subtle changes between the stages. MCI is generally diagnosed when cognitive decline is apparent, and the patient's daily living is not affected. MCI due to AD is only diagnosed when evidence of progressive cognitive decline is apparent and other causes of cognitive impairments such as depression, trauma and medical comorbidity have been ruled out (Albert *et al.*, 2011).

1.1.6.3 Dementia due to AD

A further cognitive decline from MCI can lead to dementia. Typical symptoms of dementia include; memory loss that disrupts daily activities, difficulty in planning, problem-solving, confusion with time and location, difficulty understanding visual and spatial relationships, difficulty with speaking and writing, withdrawal from social activities and changes in mood and personality (Gaugler *et al.*, 2016). Based on the severity of symptoms, the disease can be further classified as mild, moderate or severe. In the mild stage, individuals are functioning independently but may require assistance with some activities. In the moderate stage, individuals may encounter difficulty performing routine tasks, become confused, and can start exhibiting personality and behavioural changes, including suspiciousness and agitation. In the severe stage of the disease, an individual's communication becomes limited, and they are unable to perform basic activities without assistance, such as bathing (Gaugler *et al.*, 2016).

Dementia due to AD can be further categorised into "probable AD dementia" and "possible AD dementia" within the clinical setting. Probable AD dementia is primarily diagnosed when the dementia criteria have been met, other common causes of dementia have been ruled out, and the patient has a clear history of gradual cognitive decline (McKhann *et al.*, 2011). Possible AD dementia diagnosis is similar to probable AD, except a history of cognitive decline is absent, and onset is more sudden or where evidence of other causes of dementia are also present, such as cerebrovascular disease (McKhann *et al.*, 2011).

As mentioned by Bird (2018), AD can only be established at death by confirming:

- Neuropathological Aβ plaques; plaques should stain positively with Aβ antibodies and negative for prion antibodies (which are diagnostic of prion diseases).
- Intraneuronal neurofibrillary tangles; Aggregation of alpha-synuclein in the form of Lewy bodies may also be found in neurons in the amygdala; frequently, there is an accumulation of TDP-43 protein.
- Amyloid angiopathy - the numbers of plaques and tangles must exceed those found in age-matched controls without dementia.

1.2 AD brain transcriptomic perturbations

1.2.1 Gene expression

The central dogma of molecular biology is the process of encoding instructions in the DNA to RNA and then to functional products. A disruption or change in this process can lead to disease. A gene is regarded as the nucleotide sequence in DNA that provides instructions for functional products such as proteins, which is expressed through the process of transcription and translation. During transcription, the DNA of a gene serves as a template for complementary base-pairing, forming a pre-mRNA molecule which is processed into mature mRNA. The mRNA is a single-stranded copy of the gene, which then undergoes translation to form a functional product. Measuring the mRNA levels can provide a snapshot of biological activity at a molecular scale which may reflect the number of functional products such as proteins. The number of human protein-coding genes has been estimated to range between 19,901-21,306 (Salzberg, 2018).

1.2.2 Microarray technologies

Gene expression profiling is the measurement of mRNA from several genes in a sample, collectively, representing a profile of gene expression. Until the early 1990s, the traditional methods for measuring levels of mRNA involved northern blotting and reversed polymerase chain reaction (RT-PCR), a time consuming and restricted procedure. The invention of microarray's revolutionised the genomic

research by allowing global gene expression of thousands of genes to be simultaneously be measured at a relatively lower cost. Several microarray platforms have been developed for measuring gene expression, with this thesis focusing on Affymetrix and Illumina technologies due to their high replicability across platforms (Barnes *et al.*, 2005; Chen *et al.*, 2007; Consortium, 2008; Maouche *et al.*, 2008).

The basic principle of microarray technology involves immobilising thousands of pre-defined DNA spots (probes) to the surface of a plate (BeadArray). Each DNA spot contains many copies of the same probe. Complimentary sample nucleic sequence can hybridise to these DNA spots, which can be quantified by detection of fluorescent labelling. The measured fluorescence represents the abundance of a particular gene in the sample.

1.2.3 Next-generation sequencing

Next-generation RNA-sequencing (RNA-Seq) uses deep-sequencing technologies to provide a wider and comprehensive view of the whole transcriptome. The technologies have many benefit over microarray technologies, such as the ability to be free from pre-defined probes, detect sequence variations in transcribed regions, longer base pair reads (30-400bp) offer more precise mapping to the genome, as microarray short reads can be limited due to the repetitiveness of the genome, has a larger dynamic range of expression measurement and has been suggested to be more reproducible (Wang, Gerstein and Snyder, 2009).

The reproducibility between microarrays and RNA-Seq have been suggested to be very high, with the exception for transcripts with extremely high or low expression (Sîrbu *et al.*, 2012; Zhao *et al.*, 2014; Chen *et al.*, 2017). Therefore, results replicated by both microarray and RNA-Seq technologies using independent datasets can further strengthen the reliability of the results. Although RNA-Seq is superior to microarray technologies, the relatively high cost and lack of public availability of RNA-Seq data, this thesis focuses on microarray-based gene-expression data.

1.2.4 Differentially expressed genes

Many studies have undertaken gene expression profiling to identify specific genes that are significantly altered in AD brains when compared to normal aged brains. These genes are referred to as differentially expressed (DE). One major challenge with studies focusing on DE genes (DEG's) in AD is the lack of replication across independent studies. For instance, two independent AD microarray gene expression studies identified DEG's in the hippocampus brain regions. The first study (Miller *et al.*, 2013) identified 600 DEG's, while the second (Hokama *et al.*, 2014) identified 1071 DEG's, from which 105 DEG's overlapped between the two studies, however, none of which passed multiple corrections. The Miller study was based on 7 AD samples and 10 controls expression profiled on the Affymetrix platform, while

the Hakoma study contained 31 AD and 32 controls expression profiled on the Illumina platform. Replication between the Illumina and Affymetrix platform has been shown to be generally very high (Barnes *et al.*, 2005; Chen *et al.*, 2007; Consortium, 2008; Maouche *et al.*, 2008); therefore, the lack of replication between the two studies is probably down to a range of other factors including low statistical power, sampling bias and disease heterogeneity.

1.2.5 Biological pathways

Approaches such as Gene Set Enrichment Analysis (GSEA) can be used to provide functional annotation for lists of DEG's between two biological conditions. This providing an insight into biological processes or molecular functions which may be altered based on lists of perturbed genes. Numerous AD studies have highlighted disruptions in the immune response (Lambert *et al.*, 2010; Li *et al.*, 2012; Sekar *et al.*, 2015; Chen *et al.*, 2016), protein transcription/translation regulation (Li *et al.*, 2012, 2015; Liu *et al.*, 2013; Godoy *et al.*, 2014; Sekar *et al.*, 2015; Puthiyedth *et al.*, 2016) calcium signalling (Blalock *et al.*, 2011; Ramanan *et al.*, 2013; Sekar *et al.*, 2015), MAPK signalling (Miller *et al.*, 2013; Puthiyedth *et al.*, 2016), various metabolism pathways such as carbohydrates (Puthiyedth *et al.*, 2016), lipids (Paolo and Kim, 2012; Puthiyedth *et al.*, 2016), glucose (Ishii *et al.*, 1997; Liang *et al.*, 2008; Godoy *et al.*, 2014), and iron (Oshiro, Morioka and Kikuchi, 2011; Li *et al.*, 2012), chemical synapse (Blalock *et al.*, 2011; Miller *et al.*, 2013; Ramanan *et al.*, 2013), neurotransmitter pathways (Blalock *et al.*, 2011; Li *et al.*, 2012; Ramanan *et al.*, 2013) and many infectious diseases pathways (Porcellini *et al.*, 2010; De Chiara *et al.*, 2012; Kumar *et al.*, 2016). Concordance at the pathway level can sometimes replicate better than the individual gene level across studies (Maglietta *et al.*, 2010). However, these biological pathways observed to be dysregulated in AD have also been observed to be disturbed in other neurological disorders; that is, the changes observed are possibly not specific to AD. For example, the dysregulation of calcium signalling, MAPK, chemical synapse and various neurotransmitter pathways have also been implicated in PD (Edwards *et al.*, 2011; Chandrasekaran and Bonchev, 2013), glucose metabolism, and various neurotransmission pathways have been implicated in Bipolar Disease (BD) (Clelland *et al.*, 2013; H. Chen *et al.*, 2013; Khanzada, Butler and Manzardo, 2017), chemical synapse pathways have been shown to be altered in Schizophrenia (Khanzada, Butler and Manzardo, 2017; Liu *et al.*, 2017), the immune response and protein transcriptional pathways have been suggested to be altered in HD (Labadorf *et al.*, 2015), and a number of metabolism and synapse pathways have also been implicated in MDD (Zubenko *et al.*, 2014). The molecular changes in AD brains are steadily becoming apparent; however, many other neurological disorders are showing similar biological pathway disruptions.

1.3 AD biomarkers

Biomarkers can be used to detect the presence of a disease, rule out disease, determine disease risk and monitor treatment response. Currently, there is no definitive biomarker for AD, and confirmation of the disease can only be done at autopsy. Current challenges for biomarkers include (i) accurate diagnosis of AD, as a study observed 25% of patients clinically diagnosed with probable AD did not have pathological evidence of AD at autopsy (Sabbagh *et al.*, 2017), (ii) predict the rate of disease progression and conversion likelihood between the different stages of the disease, for instance, whether a patient is likely to convert from MCI to AD, and (iii) diagnosis of the disease early, ideally during its preclinical phase when initial brain changes are at a level insufficient to manifest into clinical symptoms. This may allow interventions to halt further brain damage which may have later led to dementia, which ultimately may be more effective than attempting to reverse the pathology that already exists in AD.

The primary focus for AD biomarker discovery revolved around neuroimaging and Cerebral Spinal Fluid (CSF), which have shown to be highly accurate in detecting pathophysiological and neurophysiological changes associated with AD. However, the high cost and invasiveness have shifted the field to increase research into cognitive assessments and blood-derived biomarkers that may serve as alternative biomarkers, or at the very least, be an initial screen for AD to reduce numbers of false positives (Hampel *et al.*, 2018).

1.3.1 Non-blood-based biomarkers

1.3.1.1 Neuroimaging

Neuroimaging is a process of producing images of the structure or activity of the brain or another part of the nervous system. Since AD is a disease of the brain, neuroimaging is understandably one of the obvious choices to further understand the anatomical and functional changes in AD brains. Autopsy studies have demonstrated positron emission tomography (PET) scans can accurately reflect brain Aβ levels (Ikonomovic *et al.*, 2008; Fleisher *et al.*, 2011), and as a result, can be used within the clinical environment to aid in AD diagnosis, however it cannot be used exclusively to confirm AD diagnosis. For instance, clinicians can use PET imaging to query whether an MCI patient is likely due to AD, allowing the clinician to explore alternative causes if the results are negative.

Three additional imaging biomarkers are used for research purposes, which include elevated cortical tau as a biomarker for neurofibrillary tangles, which can be measured through PET scans (Villemagne *et al.*, 2014); decreased glucose metabolism measured through fluoro-2-Deoxy-D-glucose (FDG) PET as a biomarker for decreased neuronal activity, which has been suggested to occur early in AD (Mosconi *et al.*, 2010) and brain atrophy as a biomarker for neurodegeneration, which can be measured through structural magnetic resonance imaging (MRI) (McEvoy *et al.*, 2009).

Additional neuroimaging investigations have also reported functional MRI (fMRI) can detect decreased activity in the medial prefrontal cortex of AD subjects when compared to MCI patients. While structural MRI (sMRI) can distinguish early AD brain atrophy from healthy ageing (Scheltens *et al.*, 2002), brain atrophy is associated with cognitive decline and can be used as a marker for disease progression from cognitive normal to MCI and then further to AD (Cardenas *et al.*, 2011; Del Sole, Malaspina and Magenta Biasina, 2016).

Neuroimaging has been reported to accurately diagnosis the disease early whilst also being capable of monitoring disease progression; however, the high costs associated with the technologies limits its feasibility as a widespread diagnostic tool.

1.3.1.2 Cerebral spinal fluid

CSF is a clear, colourless liquid surrounding the brain and spinal cord which supports the brain by acting as a mechanical barrier against shock, maintains pressure within the cranium and provides lubrication between surrounding bones. CSF also provides nutrients and disposes the metabolic waste from the brain. It has been well established that a decrease in CSF $A\beta_{42}$ accompanied by an increase total tau and phosphorylated tau at threonine 181 (ptau-181) levels in MCI and AD reflect neuropathology (El Kadmiri *et al.*, 2018). The high diagnostic performance of CSF as a biomarker for AD has also been confirmed at autopsy and has even been demonstrated to outperform pure clinical diagnosis (Bittner *et al.*, 2016). However, variability in Aβ measurements between batches and across clinical laboratories has hindered clinical utility. New fully automated procedures designed specifically to tackle this issue, such as electrochemiluminescence immunoassay for the quantitation of $A\beta_4$, have demonstrated stable and very precise measurements that may improve the replication accuracy of CSF as an AD biomarker (Blennow, 2017).

CSF fluid is extracted through a lumbar puncture, sometimes regarded as a relatively invasive procedure. The most common complication associated with a lumbar puncture is post-lumbar puncture headache, which has been shown to vary across different geographical sites and can be dependent on the procedure type. Although the reported incident rate is <2%, patients in the general population have been reported to fear the pain and side-effects associated with the procedure (Dekker *et al.*, 2017).

1.3.1.3 Cognitive assessments

Cognitive tests can be used as part of a series of tests, to suggest dementia in the first instance, followed by further tests to suggest the underlying disease. There are a number of different cognitive tests, with the most widely used Mini-Mental State Examination (MMSE), Clinical Dementia Rating Scale (CDR), and the AD Assessment Scale-Cognitive Subscale (ADAS-Cog).

The MMSE was first developed by Folstein and others in 1974 as a bedside screening test for dementia (Folstein, Folstein and McHugh, 1975). The test is available in many languages, making it one of the most widely used short tests to measure overall cognitive impairment in both the clinical and research setting. The test consists of a series of tasks to calculate a score based on orientation, word registration and recall, attention and calculation, language abilities and visuospatial ability. Generally, the MMSE score ranges from 0-30, where ≥25 are considered normal, 19-24 indicates early dementia, 10-19 indicates moderate dementia, and <10 is indicative of severe dementia. Scores typically decline with advancing age and increase with higher education level. The test is repeatedly used to assess cognitive decline in an individual over time, with AD subjects typically declining 3-4 points per year, and therefore can be effective at monitoring treatment effect.

The CDR is an alternative approach used to measure cognitive decline based on a clinical interview. The CDR assesses six domains to calculate a score which lies on a 5-point scale. The six domains are memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care. Patients who score 0 have no cognitive decline, 0.5 indicates questionable or very mild dementia, 1 indicates mild dementia, 2 indicates moderate dementia and 3 indicates severe dementia.

The ADAS-Cog was designed in 1984 by Rosen and others as a rating scale to assess the severity of AD based on the assessment of core symptoms associated with the disease (Rosen, Mohs and Davis, 1984). The test is primarily used in clinical trials and consists of 11 sections which assesses memory, language, praxis and orientation. The test provides a score ranging from 0-70, where ≥18 indicates greater cognitive impairment (Rockwood *et al.*, 2007). The test score has been shown to increase with age which adds complexity to the interpretation of longitudinal studies lasting several years.

In the research environment, there is no consensus for a single "gold standard" test for cognitive decline, and as a result, many studies use different cognitive tests. A study compared the ADAS-cog, MMSE and CDR scores from the same 1709 subjects and concluded the three tests were consistent in identifying cognitive decline. However, CDR and ADAS-Cog were more precise in measuring the severity of cognitive decline than MMSE (Balsis *et al.*, 2015).

1.3.2 Blood-based biomarkers

Blood-derived biomarkers offer a less-invasive, more readily accessible and potential cheaper approach for AD detection compared to CSF and some neuroimaging biomarkers. Since AD is a disorder of the brain, identifying signals in peripheral blood assumes the disease affects blood as well, or dysfunction of the blood-brain barrier (BBB) allows disease markers to leak from the brain to blood. The BBB is composed of tightly sealed endothelial cells that freely allows diffusion of oxygen and carbon dioxide. The specialised BBB endothelial transport system carries energy metabolites, nutrients, and regulatory

molecules from blood to brain, and metabolic waste products and potential neurotoxic molecules from brain to blood (Montagne, Zhao and Zlokovic, 2017). A breakdown of the BBB in AD has been detected in 20 independent human studies and has even been detected before cognitive decline is apparent (Montagne, Zhao and Zlokovic, 2017), suggesting biomarkers for AD may potentially exist in peripheral blood. Currently, there is no blood-based biomarker that is being used in the clinical setting for AD discovery; however, many methods have been proposed by individual studies, which are discussed below.

1.3.2.1 Genetics

As discussed earlier, there are a number of genetic mutations that can increase the risk of AD, with APOE genotype carrying the greatest risk. Incorporating the APOE ε4 status can improve clinical diagnostic status from 55% to 84% (Mayeux *et al.*, 1998). In addition, a study observed 50% of MCI patients who are APOE ε4 carriers, progressed to AD in 3 years compared to 20% who were APOE ε4 negative (Petersen, 2004).

Genome-Wide Association Studies (GWAS) identify genetic variants associated with a particular trait and have been performed between subjects with and without AD to identify genetic markers associated with AD. A recent study combined these AD markers into polygenic risk scores (PRS), and along with sex and age, were capable of predicting AD with 70% accuracy (70% sensitivity and 70% specificity) within a case-control dataset (3049 AD and 1554 controls) (Escott-Price *et al.*, 2015). As AD is known to share genetic risk factors with other neurological disorders, such as DLB (Orme, Guerreiro and Bras, 2018), predicting AD using only genetic variation may not be able to distinguish AD from other disorders. In addition, due to the heterogeneity of the disease, which is influenced by a combination of genetic and environmental factors, the use of DNA characteristics and phenotypic information to predict AD may not capture the full spectrum of molecular-level changes influenced by environmental factors, and as a result, may not be able to predict the onset of the disease. Therefore, genetic information should be incorporated with other biomarkers to aid AD diagnostics.

1.3.2.2 Gene expression

Several studies have attempted to exploit blood gene expression measurements for AD biomarker discovery, with initial research relying on DEG as a means to distinguish AD from non-AD subjects (Rye *et al.*, 2011; Han *et al.*, 2013). However, the limited overlap and reproducibility of DEG from independent cohorts suggests this method alone is not reliable enough (Han *et al.*, 2013). A solution to this problem would be to use information across all genes simultaneously through machine learning algorithms, which can capture minute gene expression changes, which collectively may be more informative than DEG. This technique has been demonstrated in multiple AD studies to build AD

classification models; (Fehlbaum-Beurdeley *et al.*, 2010; Booij *et al.*, 2011; Lunnon *et al.*, 2013; Roed *et al.*, 2013; Voyle *et al.*, 2016) however, as many were underpowered due to lack of samples, validation was either performed in the same subjects that were used to develop the classification model, which is known to significantly exaggerate validation performance and is referred to as an overfit, or they split their existing small dependent cohort into a training set for model development and a testing set for validation purposes, which can inadvertently lead to learning and predicting platform technical noise and batch effects rather than biology.

Previous studies have demonstrated the potential use of blood transcriptomic levels to differentiate between AD and cognitively healthy individuals; however, they are yet to be precise enough for clinical utility and are yet to be extensively evaluated on specificity by evaluating model performance in a heterogeneous ageing population of multiple diseases. This validation process is critical to determine whether the classification model is indeed disease-specific, a general indication of ill health, or an overfit.

1.3.2.3 Peptides

Peptides are short chains of amino acid monomers linked by peptide bonds. Blood Aβ peptides ratios $A\beta_{1–42}$:$A\beta_{1–40}$ have been measured by enzyme-linked immunosorbent assay (ELISA) methods, with early studies reporting little or no significant difference between AD and control subjects. However, recent studies based on ultrasensitive immunoassay techniques have reported weak significant correlations between CSF and plasma $A\beta_{1–42}$:$A\beta_{1–40}$ ratios (Hampel *et al.*, 2018). Mass-spectrometry-based studies have also reported reduced plasma $A\beta_{1–42}$:$A\beta_{1–40}$ ratios in patients with AD when compared to controls (Pannee *et al.*, 2014). In addition, immunoprecipitation coupled with mass-spectrometry identified $APP_{699-711}$/$A\beta_{1–42c}$ and $A\beta_{1–42}$:$A\beta_{1–40}$ ratios were able to predict Aβ-positive individuals based on PET scans with 90% accuracy (Nakamura *et al.*, 2018), further suggesting plasma Aβ measurements may be useful as biomarkers for AD. However, other disorders including cardiovascular and cerebrovascular have also reported altered plasma Aβ-levels (Roeben *et al.*, 2016; Hilal *et al.*, 2017), suggesting further investigations are required to determine the specificity of this biomarker.

1.3.2.4 Proteins

Tau protein is an obvious choice to measure as an AD biomarker, however, reported plasma levels been contradictory, with individual studies reporting AD subjects having no significant change, mild elevation, significant elevation or even reduced levels of plasma Tau protein (Mattsson *et al.*, 2016). In addition, a longitudinal AD study reported increased plasma tau levels were correlated with future cognitive decline, increasing atrophy and hypermetabolism (Mattsson *et al.*, 2016). However, subjects with Frontotemporal lobar degeneration (FTLD) have also been observed to have elevated plasma tau levels

(Foiani *et al.*, 2018); therefore, plasma tau levels alone may not be able to distinguish AD from other disorders that have tau pathology.

An additional 163 candidate blood-derived proteins were identified as potential AD biomarkers from 21 separate studies (Kiddle *et al.*, 2014). The overlap of discovered proteins between studies was limited, with only four biomarkers α-1-antitrypsin, α-2-macroglobulin, apolipoprotein E and complement C3 found to replicate in five independent cohorts. However, a follow-on study discovered these four biomarkers were not specific to AD and were also discovered to be associated with other brain disorders including Parkinson's Disease (PD) and Schizophrenia (SCZ) (Chiam *et al.*, 2014).

1.3.2.5 Metabolites

Metabolites are small intermediate products formed from the metabolic reactions caused by various enzymes and have been found to be perturbed in AD brains (Inoue *et al.*, 2013). The blood consists of thousands of small metabolites at any given time, which may have transported from the brain through the BBB, providing a potential signal of the brains activity. Using ultra-performance liquid chromatography (UPLC) coupled with mass spectrometry (UPLC-MS), a multicentred study compared 495 plasma metabolites to identify a 7-metabolites profile capable of differentiating AD and aMCI from healthy controls. The 7 metabolites included; three amino acids (glutamic acid, alanine, and aspartic acid), one non-esterified fatty acid (22:6n-3, DHA), one bile acid (deoxycholic acid), one phosphatidylethanolamine [PE(36:4)], and one sphingomyelin [SM(39:1)] (Olazarán *et al.*, 2015). These metabolite biomarkers still require further validation in similar disorders.

1.3.3 Limitations

The current time-consuming clinical diagnosis of AD is based on cognitive tests and neuroimaging, which has been shown to have an estimated accuracy of 75% at autopsy (Sabbagh *et al.*, 2017). Clearly, a more accurate, easily accessible, relatively cheap, and quicker method for AD diagnosis is required. CSF biomarkers seem the most promising; however, as a relatively intrusive method, it may not appeal to the broader population. Blood-based biomarkers are more appealing, with many individual studies consistently demonstrating the ability of various biomarkers capable of differentiating AD from healthy controls; however, many lack reproducibility and specificity are rarely tested. Blood-based biomarkers require an extensive evaluation with age-related and other similar disorders to be useful as a clinical screening test for the general population.

1.4 AD treatment

To date, no disease-modifying treatment is available for AD. The underlying cause of the disease is currently unknown, making it extremely difficult to develop treatments to cure, reverse or even halt the neuropathological process of AD. In addition, effective treatment response in animal models does

not necessarily reflect the outcomes in humans. The slow clinical recruitment of patients and the relatively long time needed to observe the effectiveness of treatment on disease progression further contribute to the difficulty in developing an effective treatment.

1.4.1 Pharmacological

Symptomatic relief is available in the form of cholinesterase antagonist's donepezil, rivastigmine and galantamine, and an N-methyl-D-aspartate (NMDA) inhibitor, memantine. These drugs only provide temporary symptomatic relief, and do not modify the underlying disease pathology, and may not have the same effect on all patients (Calciano *et al.*, 2010; Anand, Gill and Mahdi, 2014).

The three cholinergic inhibiting compounds all have slightly different pharmacological properties, but all work by preventing the enzyme acetylcholinesterase breaking down acetylcholine, a neurotransmitter associated with memory. The three cholinesterase inhibitors are effective for mild to moderate AD and have been reported to temporarily improve a patients ADAS-cog score by an average of 2.7 points (Birks, 2006).

In AD, damage to brain cells releases excessive neurotransmitters such as glutamate. In normal conditions, glutamate plays an essential role in learning and memory. However, excessive glutamate causes overexcitation of nerve cells, which can lead to cell damage and/or death. Memantine works by blocking glutamate receptors NMDA, which prevents the excessive activity of glutamate and temporarily stabilising cognitive decline. Memantine has been shown to be less effective than cholinergic inhibitors, but in combination, have proved to be more effective when compared to standalone treatment with each drug (Parsons *et al.*, 2013). The side-effects associated with these drugs are generally mild and include gastrointestinal side-effects such as diarrhoea and nausea (Loi *et al.*, 2018).

1.4.2 Non-Pharmacological

Non-pharmacological treatment is often used to maintain or improve cognitive function without the use of medication, and like pharmacological treatments, do not slow, stop or reverse the neuropathological characteristics of AD. These treatments typically involve computerised cognitive training, listening to music or physical exercise to improve overall cognitive function ('2018 Alzheimer's disease facts and figures includes a special report on the financial and personal benefits of early diagnosis', 2018). Non-pharmacological treatment response in AD has been inconsistent, with very little evidence to support a specific treatment for AD. Nevertheless, individuals have reported improvement from these methods, and these methods may be more suited to be personalised per patient, although this may be more resource-intensive (Loi *et al.*, 2018).

1.4.3 Future treatments

Changes in AD brains occurs several years in advance before the onset of clinical symptoms, identification of patients at this preclinical stage may allow interventions to halt further brain damage, which may have later led to dementia. This may be more effective than attempting to reverse the pathology that already exists in AD patients. However, the major drawback here is the lack of diagnostic methods to identify AD early, and therefore, the majority of treatments are focusing on clinically diagnosed AD patients.

Although the exact cause of the disease is unknown, clinical trials are primarily targeting specific processes which may reverse or prevent further accumulation of Aβ and NFT. Several clinical trials are currently ongoing and are primarily targeting Aβ clearance, BACE inhibition, γ-secretase inhibition, stabilising Tau, inhibiting Tau aggregation or P-Tau clearance, however, from 1984-2018, more than 200 of these drugs have entered phase 2 of a clinical trial but none have made it to clinical use (Loi *et al.*, 2018).

1.4.3.1 Drug repositioning

The typical time for an AD drug development program takes 13 years with costs estimated to be £4.4 billion, which is considerably higher than the estimated £624.1 million for cancer treatment development (Cummings, Reiber and Kumar, 2018). Drug repositioning can provide an alternative strategy, where established drugs are repurposed for new therapeutic targets, reducing the costs and time associated with drug development, whilst making the process accessible to academic centres along with pharmaceutical and biotechnology companies.

Drug repositioning has been successfully applied to a variety of diseases, including cancer, cardiovascular disease, stress incontinence, irritable bowel syndrome, obesity, erectile dysfunction, smoking cessation, psychosis, attention deficit disorder and Parkinson's disease (Corbett *et al.*, 2012). With the wealth of open publicly available biological data, drug repositioning can be applied to identify possible treatments for AD as well.

1.5 Neurological disorders

Neurological disorders are diseases of the nervous system, which may impair the brain, spinal cord, peripheral nerve or neuromuscular function. There are more than 600 different neurological disorders, with AD accounting for just one. Many neurological disorders show similar symptoms, neuropathology or brain changes at a molecular level to AD. Some of these disorders are incorporated into this thesis to aid in identifying AD-specific brain and blood changes. Therefore, a brief overview of each disorder and similarities to AD are discussed in the following sections.

1.5.1 Dementia

As discussed earlier, dementia is a broad term given to individuals who exhibit greater cognitive decline than expected in normal ageing. AD accounts for an estimated 60-80% of reported dementia cases, however, upon autopsy, an estimated half of these individuals have shown to involve additional pathology consistent with other dementias (Gaugler *et al.*, 2016).

1.5.1.1 Dementia with Lewy bodies

DLB is a neurodegenerative disease accounting for 10-15% of all dementias, with an incidence rate estimated to be 6 in every 100,000 (Savica *et al.*, 2013) and age of onset typically between the age of 50 to 83 years. The exact cause of the disease is unknown, but the disease shares risk loci with AD in the APOE E4 allele, and with PD, in variants SNCA (chromosome 4) and GBA (chromosome 1) (Orme, Guerreiro and Bras, 2018).

The neuropathological hallmark of the disease includes abnormal clumps of protein alpha-synuclein (Lewy bodies) in neurons, which is accompanied in most cases by Aβ deposition and occasional NFT. The accumulation of these proteins disrupts normal cognitive functions in brain regions such as the cerebral cortex, areas of the temporal lobe, such as the superior temporal gyrus, parahippocampal gyrus and the limbic system areas, including the hippocampus and amygdala (Armstrong, 2012). Disruption to these brain regions causes symptoms to initial include sleep disturbances and hallucinations, which can occur in the absence of memory impairment. This is typically followed by progressive cognitive decline and parkinsonism (a clinical syndrome characterised by tremor, bradykinesia, rigidity and postural instability).

In addition to sharing the same risk loci APOE, many similarities exist between DLB and AD, from symptoms of progressive cognitive decline and dementia to the hallmark neuropathology, which includes Aβ and NFT. A study found no significant difference in cognitive decline over 12 months for AD and DLB subjects when baseline age, gender and CDR rating are matched (Walker *et al.*, 2012); however, DLB subjects have reported lower scores in visual memory, visuoperceptive, and visuoconstructive tests when compared to AD (Noe *et al.*, 2003).

1.5.1.2 Frontotemporal lobar degeneration

Frontotemporal lobar degeneration (FTLD) is a heterogeneous group of disorders which is characterised by the progressive neurodegeneration of the prefrontal and temporal cortices. FTD is the most common FTLD and accounts for 5-15% of dementia cases (Mohandas and Rajmohan, 2009). The prevalent genetic risk factors for FTD include PGRN (progranulin) and MAPT (microtubule-associated protein tau), both located on chromosome 17.

The neuropathological hallmark of the disease can be categorised into three broad groups; frontotemporal lobar degeneration with tau, TDP-43 or FET protein accumulation (Mackenzie and Neumann, 2016). The initial symptoms include personality and behavioural changes, which is dissimilar to the memory issues initially observed in AD. Similar to AD, FTD can occur above the age of 65 years of age; however, over 60% of reported cases occur between 45-60 years of age (Gaugler *et al.*, 2016).

1.5.1.3 Parkinson's Disease

PD is a progressive disorder of the nervous system that affects several regions of the brain devoted to the control of balance, movement, emotions and general cognition. PD is the second most common neurodegenerative disorder, affecting 5 to >35 per 100,000 depending on demographics (Poewe *et al.*, 2017). Typically the onset of PD occurs after the age of 50, with approximately 30% developing dementia (Aarsland, Zaccai and Brayne, 2005). A number of genes have been associated with PD, including LRRK2 (chromosome 12), GBA (chromosome 1), and SNCA (chromosome 4).

The disease is characterised by neuronal loss and abnormal aggregates of alpha-synuclein protein in the substantia nigra brain region. Similar to dementia with Lewy body, the alpha-synuclein contribute to structures called Lewy body that displaces other cell components. Lewy bodies initially occur in the cholinergic and monoaminergic brainstem neurons but are also found in limbic and neocortical brain regions with disease progression (Poewe *et al.*, 2017).

PD and AD are distinct disorders that share common features. Both are progressively slow neurodegenerative diseases that typically begin late in life, although initial symptoms differ, both diseases can lead to dementia. The neuronal degeneration includes intraneuronal inclusions and the hallmark pathology of PD, alpha-synuclein, has also been reported in AD brains, but are concentrated mainly in limbic brain regions (Poewe *et al.*, 2017).

1.5.1.4 Creutzfeldt-Jakob disease

Creutzfeldt-Jakob disease (CJD) is an extremely rare and fatal disorder characterised by misfolding proteins (prion) that cause other proteins to misfold and malfunction throughout the brain. CJD develops in approximately 1 in every million, with the onset of symptoms around 70 years of age and mean survival time of only 6 months. Over 90% of subjects died within a year from the onset of symptoms (Mackenzie and Will, 2017).

There are two types of CJD; sporadic (sCJD) and variant (vCJD). sCJD is the most common form of the disease, where the cause of the disease is unknown. In contrast, vCJD is caused by the consumption of meat, where the animal was infected by bovine spongiform encephalopathy. The initial symptoms in both forms of the disease include problems with muscular coordination; personality changes, including

impaired memory, judgment, thinking; and impaired vision, and eventually lead to dementia. CJD subject's cognitive decline is more rapid than AD.

1.5.1.5 Vascular dementia

VaD is caused by blockage or damage to blood vessels, which causes strokes or bleeding to the brain. The location of the brain damage determines the impact on the individual's physical and mental functions, including if dementia occurs. Initial symptoms often include impaired judgement and lack of organisation, which contrasts memory issues associated with initial symptoms of AD.

Diagnosis of VaD is extremely difficult, with clinical diagnosis achieving an estimated 50% sensitivity at autopsy (McAleese *et al.*, 2016). Pure VaD is suggested when neuropathology vascular lesions are present without neurodegenerative pathologies such as AD or Lewy body pathology. However, in 40% of dementia cases, patients exhibit deposition of Aβ and NFT consistent with AD, which can be regarded as mixed dementia (Gaugler *et al.*, 2016).

1.5.1.6 Mixed dementia

Mixed dementia is a term given to an individual who exhibits hallmark abnormalities associated with more than one cause of dementia. In most cases, AD and vascular dementia, also known as Mixed Vascular-Alzheimer Dementia (MVAD), are most observed to coexist, followed by AD, vascular dementia and DLB (Gaugler *et al.*, 2016).

1.5.2 Amyotrophic lateral sclerosis

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease that affects brain nerve cells and the spinal cord. ALS affects approximately 2 in 100,000 (*The ALS Association*, no date), with the onset of symptoms typically occurring between the age of 50-65. The exact cause of the disease is unknown; however, 19 genes have been associated with the disease. Mutations in superoxide dismutase 1 (SOD1) located on chromosome 21, accounts for approximately 20% of the inherited version of ALS (Zarei *et al.*, 2015).

In normal motor function, messages from the brain motor neurons are transmitted to spinal cord motor neurons which forwards these messages to muscles. In ALS, these motor neurons degenerate, causing a disruption to the signalling network to muscle. The muscles are unable to function correctly, leading to symptoms that include muscle weakness, twitching and cramping, which eventually cause muscle impairment. Severe cases of ALS will develop symptoms of dyspnea and dysphagia (Zarei *et al.*, 2015).

The neuropathology of ALS involves an abnormal accumulation of proteins in the cytoplasm of neurons, which causes the neurons to dysfunction. These proteins include ubiquitinated TAR DNA-binding protein (TDP-43) in the frontal cortex, temporal cortex, hippocampus and striatum, which has also been

observed to occur in FTD (Saberi *et al.*, 2015). Dementia in ALS is uncommon; however, degeneration of the frontotemporal cortex can lead to mild cognitive impairment, specifically affecting verbal praxis and fluency (Strong, Kesavapany and Pant, 2005).

When dementia is absent in ALS, the clinical symptoms of AD and ALS are unique to each disorder. Nevertheless, studies have identified similarities in ALS and AD subjects, including the activation of IRE1α-XBP1 and ATF6 pathways of the unfolded protein response (Montibeller and de Belleroche, 2018), and multiple rare reports of NFT in ALS brains (Guiroy *et al.*, 1993; Kokubo and Kuzuhara, 2004; Coan and Mitchell, 2015; Takeda, 2018)

1.5.3 Bipolar Disorder

BD, previously known as manic depression, is a mood disorder characterised by extreme fluctuation in an individual's mood and depression. The disorder affects approximately 1 in 100 people in their lifetime with a typical onset of disease between the age of 15-19 but can occur at any age (*Bipolar disorder - NHS*, no date). The exact cause of the disease is unknown; however, physical, environmental and social factors have been suggested to contribute to the increased risk of developing the disease. GWAS studies have identified 18 loci associated with BD, with many replicating in independent studies (Vieta *et al.*, 2018).

The brain changes involved in the disease include dendritic spine loss in the prefrontal cortex (Vieta *et al.*, 2018), alterations in neuronal interconnectivity, atrophy of the hippocampus and reduced cortical thickness (Harrison, Colbourne and Harrison, 2018). These brain changes lead to cognitive dysfunction, increased cognitive decline in older patients (Lewandowski *et al.*, 2014), and dementia in long-term cases (Forlenza and Aprahamian, 2013).

BD is usually distinguishable from AD due to the early age of onset; however, BD can occur over the age of 50, which can make behavioural impairment challenging to differentiate from AD. In addition, the following symptoms have been reported in both disorders; agitation, euphoria, disinhibition overactivity without agitation, aggression, dysphoria, apathy, impaired self-regulation, and psychosis (Besga *et al.*, 2015). Furthermore, a study identified altered epigenetic regulation related to neuroinflammation, synaptic integrity, and neuroprotection in both AD and BD (Rao *et al.*, 2012)

1.5.4 Huntington's disease

Huntington's disease (HD) is an autosomal disease with aspects of cognitive decline, which can progress to dementia in severe and rare cases. Therefore, the disease is generally classed as a motor disorder rather than dementia. The onset of HD is typically around the age of 40, with a prevalence of approximately 5-10 in every 100,000 (Reiner, Dragatsis and Dietrich, 2011).

HD individuals have an expanded CAG repeat in the huntingtin gene located on chromosome 4. The resulting huntingtin protein contains excessive glutamine with additional hydrogen bonds that cause the proteins to be "sticky" and become entangled. The aggregated proteins accumulate and interfere with nerve cell functions. Protein aggregates primarily accumulate in the basal ganglia, but can also affect the temporal and frontal lobes (Montoya *et al.*, 2006). Typical symptoms include movement issues, such as involuntary movement, cognitive issues, such as difficulty organising, and psychiatric issues such as insomnia.

Similar to AD, HD involves the clumping of proteins that disrupt the neuronal network in different regions of the brain, impairing functions associated with the normal function. This leads to the manifestation of memory issues in AD, while movement is primarily affected in HD.

1.5.5 Major Depressive Disorder

Major depressive disorder (MDD), clinically referred to as depression, is a debilitating disease characterised by mood swings, diminished interest, impaired cognitive function and vegetative symptoms, such as disturbed sleep or appetite (Otte *et al.*, 2016). The disease does not follow Mendel's law of inheritance and is caused by a combination of many genes and environmental factors (Kiyohara and Yoshimasu, 2009).

Both MDD and BD patients exhibit depressive symptoms; however, BD patients also have episodes of mania, which makes both disorders clinically distinguishable from one another. MDD symptoms can initially be difficult to distinguish from AD, and both diseases are associated with hippocampal and frontal atrophy.

1.5.6 Multiple Sclerosis

Multiple sclerosis (MS) is a chronic autoimmune disease affecting an approximate 289 per 100,000 individuals (Mackenzie *et al.*, 2014). The typical onset of the disease is around 20-40 years of age, with diagnosis after the age of 50 rare (Polliack, Barak and Achiron, 2001). The exact cause of the disease is unknown; however, the literature suggests a combination of genetics and nongenetic triggers, such as a virus, metabolism, or environmental factor thought to initiate the disease that repeatedly attacks and damages the myelinated axons of the CNS (Goldenberg, 2012). This interferes with the brain communicating with parts of the body, resulting in symptoms such as loss of motor function, speech, loss of coordination, and cognitive impairment. In extremely rare cases, MS can lead to dementia, and therefore is generally not classed as dementia.

The hallmark pathology of MS involves focal demyelinated plaques in the CNS, with degrees of inflammation, gliosis, and neurodegeneration (Popescu, Pirko and Lucchinetti, 2013). MS is very distinct

from AD, in that MS is an autoimmune disease with earlier symptomatic onset. Both diseases exhibit cognitive decline, with AD more severe. Chronic inflammation with microglia activation has been suggested to play a vital role in both AD and MS pathogenesis; however, hallmark pathology for both diseases is distinctly different (Dal Bianco *et al.*, 2008)

1.5.7 Schizophrenia

Schizophrenia is a complex syndrome with many symptoms. The disease prevalence is approximately 32 per 100,000 (Reilly *et al.*, 2012), with a typical onset of symptoms in the early twenties (Gogtay *et al.*, 2011). An estimated 15% of schizophrenia are late-onset cases where disease onset is around 40 years of age, while a rarer 4% of cases represent a very late-onset version of the disease with onset around 60 years of age (Howard, 2000).

Similar to most of the disorders discussed previously, the exact cause of the disease is unknown, with a combination of genetics and environmental factors suspected to be the cause. Neuropathological studies suggest the disease is a functional disease without organic factors. Brain abnormalities primarily involve volume change, reduced neurons, size of neurons and change in neuronal arrangement across the entire brain (Iritani, 2007). Although the exact mechanism of the disease is unknown, most theories revolve around the deficiency or excess of neurotransmitters (dopamine, serotine, and glutamate) and neurochemicals (aspartate, glycine and GABA) (Patel *et al.*, 2014).

Brain changes in the disease lead to cognitive decline and change the way an individual thinks and behaves. The symptoms of the disease can be classed as “positive”, where changes relate to behaviour and thought, such as hallucinations and delusions, or “negative”, where individual become withdrawn and express lack of rational functions, such as appear emotionless (*Schizophrenia - Symptoms - NHS*, 2016). Patients with schizophrenia may not develop dementia over time, but are at greater risk (Siavelis *et al.*, 2016a)

AD and schizophrenia are distinct diseases with the typical onset and symptom distinguishable from one another; nevertheless, genetic variations associated with endosomal trafficking, autophagy, and calcium channel signalling are shared between the two diseases (DeMichele-Sweet *et al.*, 2018).

1.5.8 Summary

With over 600 different neurological disorders, it’s no surprise AD shares similar genetic variation, neuropathology, structural brain changes and clinical symptoms with other diseases. The similarities of brain changes occurring in the neurological disorders discussed in this thesis are summarised in Table 1.1. This further complicates clinical diagnosis, which is evident by a large number of misdiagnosed AD

patients identified at autopsy (Gaugler *et al.*, 2016). A clear need to identify AD-specific changes are required, which may aid in accurate AD diagnosis and therefore improve the chances of drug trials.

Table 1.1: Similarities of brain changes across neurological disorders

Brain changes	AD	DLD	FTLD	PD	CJD	VD	MD	ALS	BD	HD	MDD	MS	SCZ
Brain Atrophy	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Neurodegeneration	✓	✓	✓	✓	✓			✓				✓*	
Cognitive Decline	✓	✓	✓	✓	✓	✓*	✓*	✓	✓	✓	✓*	✓	✓
Dementia	✓	✓	✓	✓	✓	✓*	✓*		✓*	✓*	✓*	✓*	✓*
Aβ	✓	✓					✓*						
NFT	✓	✓*	✓				✓*	✓*					
Lewy Body		✓		✓			✓*						
TDP-43 protein			✓					✓					
FET protein			✓										
Altered Prion					✓								
Vascular Lesions						✓	✓*						
Huntingtin Protein										✓			

The table illustrates the overlap of brain changes across the neurological disorders discussed in this thesis. Brain changes with a star (*) occur in rare cases. Abbreviations: AD: Alzheimer's Disease, DLD: Dementia with Lewy Bodies, FTLD: Frontotemporal Lobar Degeneration, PD: Parkinson's Disease, CJD: Creutzfeldt-Jakob Disease, VD: Vascular Disease, MD: Mixed Dementia, ALS: Amyotrophic Lateral Sclerosis, SCZ: Schizophrenia, NFT: Neurofibrillary Tangles and TDP-43: TAR DNA-binding protein 43

1.6 Conclusions

The increase in life expectancy has resulted in a rise in age-related disorders, including AD. The most common form of dementia is AD, which was first described over a century ago, however, to date, there still exists a lack of understanding molecular changes specific to the disease, a clinically established robust blood-based biomarker for accurate diagnosis and a lack of therapeutic treatments.

1.7 Thesis overview

This thesis analyses novel and existing transcriptomic data to further understand the genetic and biological mechanisms associated with AD. The investigations are presented in four analysis chapters consisting of three peer-reviewed publications. The different cohorts used in this thesis are described within individual chapters.

1.7.1 Transcriptomic Analysis of Probable Asymptomatic AD and AD Brains

The thesis begins by investigating the first human brain microarray gene expression profiling of AsymAD subjects who were clinically free from dementia; however, upon autopsy were discovered to be consistent with mild AD pathology. Standard transcriptomic analysis, coupled with a system-biology approach, was used to identify biological perturbations in AsymAD brains, which may reveal mechanisms for their AD vulnerability. Subsequently, this chapter has been peer-reviewed and published in the *Brain, Behaviour, and Immunity* Journal.

1.7.2 A Meta-analysis of Alzheimer's Disease Brain Transcriptomic Data

This chapter combines existing and novel AD brain microarray gene expression studies in the largest known AD meta-analysis to date. In addition, non-AD neurological disorders (PD, BD, Schizophrenia, MDD and HD) were incorporated into the analysis to identify specific molecular changes associated with AD brains. Furthermore, both brain regions spared and affected by hallmark AD pathology were analysed to reveal transcriptomic changes and disease mechanisms most likely involved with the accumulation of tangles and plaques. This chapter is presented as a peer-reviewed publication in the *Journal of Alzheimer's Disease*.

1.7.3 Automated Drug-repositioning from a disease expression signature

This chapter develops an automated drug repositioning pipeline to query any disease gene expression signature in the reprocessed connectivity map (CMap) database using novel algorithms to identify suitable small compounds that may potentially intervene with the disease of interest with positive effects. This drug repositioning pipeline was applied to the two AD expression disease profiles identified in the first two chapters to find compounds that may intervene with the disease during the asymptomatic and symptomatic stage.

1.7.4 Working Towards a Blood-Derived Gene Expression Biomarker Specific for Alzheimer's Disease

In this final analysis chapter, a machine learning approach is used to identify a blood-based gene expression biomarker, which could discriminate AD from cognitive healthy and age-related diseased subjects. The expression signature is further validated in an independent cohort.

1.8 Details of methodologies

Since this thesis incorporates a series of publications, an overview of methods is detailed within individual chapters, including any applications and programming packages used during the analysis. However, details of more complex methods and justification behind selecting specific data processing and analysis methods are provided below.

1.8.1 Microarray Gene Expression processing

This thesis analysis both novel and existing microarray gene expression datasets. Various microarray platforms exist, with this thesis only incorporating data generated on the Illumina and Affymetrix platform as replication between the two has been demonstrated to be very high (Barnes *et al.*, 2005; Chen *et al.*, 2007; Consortium, 2008; Maouche *et al.*, 2008). No universal data processing or analysis pipelines exist for microarray data; therefore, a data processing pipeline was developed using literature recommended techniques which best suited the data in this thesis. The following is the overall sequence of Quality Control (QC) procedures performed on the datasets, with further details on specific applications used to perform the analysis provided in individual chapters.

1.8.1.1 Affymetrix platform background correction

Microarray's measure fluorescence intensities, which are subject to a range of different sources of noise, both between and within arrays (Silver, Ritchie and Smyth, 2009). Background correction is an important procedure which attempts to address this variation. Affymetrix platforms consist of GeneChips with 25-base perfect match (PM) probes accompanied by a mismatch (MM) probe differing at the complimentary 13th position (Rouchka, Phatak and Singh, 2008). The probe design hypothesised MM probes would represent background noise, which could be subtracted from the PM intensity to determine the true level of gene expression. MAS5.0 is a background correcting method that exploits this probe design by calculating a robust average of the logged PM-MM values for all probes.

An alternative method for background correcting Affymetrix platform generated data is gcRMA. This method uses estimators derived from statistical models that use probe sequence information to calculate background intensity (Wu *et al.*, 2004). A study compared 9 different background correction methods, including MAS5.0 and gcRMA, to RT-PCR as a reference and concluded a difference in performance between most of the algorithms was not statistically significant and noted: "MAS5.0 performed well in their study" (Gyorffy *et al.*, 2009). In contrast, another study reported gcRMA performs significantly better than MAS5.0 (Harr and Schlötterer, 2006) as studies have observed higher MM values when compared to PM values, implying MM values do not accurately represent background noise (Rouchka, Phatak and Singh, 2008).

The MAS5.0 background correction method was one of the first to be available to pre-process Affymetrix platform generated data and naturally was widely adopted. As discovered in the "A Meta-analysis of Alzheimer's Disease Brain Transcriptomic Data" chapter, many publicly available microarray datasets were available in a pre-processed form, where raw data was already MAS5.0 background corrected. To maintain consistency, raw Affymetrix datasets in this chapter were MAS5.0 background corrected.

The "Automated Drug-repositioning of Transcriptomic Disease Signature" chapter focuses on DEG. The gcRMA method has been suggested to best suited for the detection of DEG (Harr and Schlötterer, 2006) and since the CMap data was available in its raw intensity format, gcRMA was the chosen background correcting method for this analysis.

1.8.1.2 Illumina platform background correction

Illumina microarray probe sequences are typically 50 nucleotide sequences, with 30 replicates on each array accompanied by over a 1000 negative non-specific control sequences. These control probes do not correspond to any expressed genomic sequence and serve as the background noise measure. Similar to Affymetrix platform, numerous background correction methods have been proposed to account for intensity variation, with Model-Based Background Correction (MBCB) shown to be more precise for gene expression measurements (Allen, Chen and Xie, 2010). The MBCB uses the negative control expression values to develop a convolutional model to estimate the background noise and subsequently, expression values for each probe (Ding *et al.*, 2008). The MBCB algorithm was applied to Illumina generated datasets when possible.

1.8.1.3 Normalisation

Data normalisation is performed to account for variation between samples and experimental batches. Many normalisation methods have been developed for microarray data, with Robust Spline Normalisation (RSN) suggested as being one of the most accurate (Schmid *et al.*, 2010). The RSN algorithm combines the features of quantile and loess normalisation, and as suggested by Schmid *et al.*, (2010), was applied after background correction and log2 transformation of the raw microarray data.

1.8.1.4 Detection of sex markers

Sex was predicted using the R package Microarray Sample Sex Identifier (massiR) R package, with discrepancies between predicted and recorded sex removed from further analysis. The massiR package uses the expression values of uniquely mapping Y chromosome probes from the data to calculate the expression variance across all samples. Samples are then partitioned into two clusters, where the cluster with the highest Y-chromosome probe values are classified as males and the remaining classified as females (Buckberry *et al.*, 2014).

1.8.1.5 Detection of expressed probes

Microarray's consists of thousands of probes, each providing a measured expression intensity. Biologically, many are likely to be associated with "unexpressed genes" and simply represent noise due to non-specific binding. Filtering these "unexpressed genes" can eliminate noise (Lazar *et al.*, 2013) and increase differential power (Hackstadt and Hess, 2009). Studies have incorporated arbitrary

expression cut-off values to determine whether a gene is expressed (Hackstadt and Hess, 2009); however, this thesis takes a more biological approach and uses the log2 expression values of nine literature defined quality control genes (QCg) to determine the cut-off values in individual datasets. Five QCg's are Y chromosome genes (KDM5D, PRKY, RPS4Y1, USP9Y, UTY) which are only expressed in male tissues (Vawter *et al.*, 2004), one QCg is an X chromosome gene (XIST) which is only expressed in female tissue (Vawter *et al.*, 2004), and the remaining three QCg's (ADIPOR1, BNIP3L, MKRN1) are expressed in all normal cells and tissues including multiple regions of the brain (Chang *et al.*, 2011).

Initially, a gene with an expression value above the 90th percentile in over 80% of samples was deemed as "expressed", with the remaining "unexpressed" probes removed. An expression boxplot of the nine QCg's is created by sex and used to verify each QCg is correctly classified as expressed or unexpressed in the correct sex groups, adjusting the filtering percentile if required.

1.8.1.6 Correcting for unwanted variation

Several individual processes are involved during sample preparation and during microarray gene expression, including, transcriptional steps, labelling, hybridisation and intensity measurement. Each of these steps can introduce technical variations (Richard *et al.*, 2014). Gene expression microarrays are also affected by non-biological variations, such as reagent lot numbers, different technicians and even atmospheric ozone levels (Chen *et al.*, 2011). In addition to measurable variables, gene expression data are affected by factors that are unknown, unmeasured or too complicated to capture through simple models (Leek and Storey, 2007). These variations are referred to as "batch effects", and should be addressed during microarray data processing as they have been demonstrated to have a greater effect on the data than the underlying biological signal (Leek *et al.*, 2010).

It is impossible to measure and account for all variables that are influencing how genes are expressed, however, a study has demonstrated through a technique called Surrogate Variable Analysis (SVA), that one can predict and account for these latent effects (Leek and Storey, 2007). SVA uses expression values across genes to estimate the large-scale effects of all unmodelled factors. Known biological variations can be provided to SVA, such as diagnosis and sex, which are excluded from the model (Leek *et al.*, 2012). Since this thesis analysed publicly available gene expression datasets, with many lacking detailed phenotypic information required for robust batch effect removal, SVA was used to identify latent variables. Any significant variation identified by SVA was then adjusted out using an Empirical Bayes method referred to as "Combating Batch Effects When Combining Batches of Gene Expression Microarray Data" (COMBAT) (Chen *et al.*, 2011).

1.8.1.7 Outlier sample detection

A network-based approach was used to identify outlier samples as described by Oldham, Langfelder and Horvath (2012). Within each diagnosis and gender group, sample relationship was calculated based on signed weighted correlations. Sample similarity measures are calculated and used to construct a sample network, which includes standardised sample connectivity (Z.K), standardised sample clustering coefficient (Z.C) and mean inter-sample adjacency (ISA). As recommended by Oldham, Langfelder and Horvath (2012), samples with a Z.K value below -2 were deemed outliers, samples removed, sample networks reconstructed, and sample similarity measures re-examined. If necessary, this process is repeated until no further sample is identified as an outlier.

1.8.1.8 Annotating platform-specific probe ID's

This thesis uses gene expression data generated from the Illumina and Affymetrix microarray platforms, both of which use unique platform-specific probe identifiers. Therefore, probe identifiers were converted to their corresponding unique Entrez Gene Identifiers (Entrez ID) to allow information to be comparable across platforms and datasets. Entrez ID represents a unique stable identifier for a specific gene (Maglott *et al.*, 2011) . Multiple probes can represent a single Entrez gene ID, from which the probe with the highest expression across all samples was retained.

1.8.1.9 PCA

Principal component analysis (PCA) was performed to calculate the variation in the gene expression data. PCA performs linear transformations to reduce the high dimensionality of the data, whilst retaining as much variation as possible. This creates principal components (PC's), which are uncorrelated and ordered by variance explained. PC's were graphically visualised against one another in a 2-dimensional plot. Clustering samples indicated similarities between the samples, with multiple clusters that are not explained by known biological variation an indication of batch effects (Ringnér, 2008). PCA plots were generated and examined throughout the QC process to ensure additional variation was not being introduced into the data after every data processing step.

1.8.1.10 Dataset compatibility

During the "A Meta-analysis of Alzheimer's Disease Brain Transcriptomic Data" chapter, multiple publicly available datasets from different microarray platforms, BeadArrays/GeneChips, tissue source and diseases were integrated for biological interpretation. To ensure dataset compatible, the gene expression datasets underwent additional QC based on six quantitative measures to identify problematic datasets. As described by Kang *et al.*, (2012), the six quantitative measures are:

1. Internal Quality Control Index (IQC): This metric compares pair-wise differences between studies to identify outlier studies from quantified co-expression dissimilarities. The analysis only

uses information directly from the expression data and does not use any external information. A study with a small IQC indicates heterogeneous co-expression structure with other studies and should be considered as a candidate problematic study for exclusion.

2. External Quality Control Index (EQC): This metric is similar to the IQC but uses external pathway information. A study with a small EQC value is indicated to be lowly associated with the remaining studies in terms of gene pairwise correlation structure and should be considered as a candidate problematic study.
3. Accuracy Quality Control Gene Index (AQCg): This metric assesses the reproducibility of DEG of a study compared to a meta-analysis of all remaining studies. A small AQCg score indicates the DEG's in a study was not reproducible when compared to DEGs detected by the meta-analysis.
4. Accuracy Quality Control Pathway Index (AQCp) – This metric is similar to AQCG but uses enrichment pathways instead of DEG when assessing reproducibility across studies.
5. Consistency Quality Control Gene Index (CQCg): This metric is similar to AQCg but uses a DE ranking system. The ranked DEG from individual study analysis and meta-analysis are compared using Spearman's rank correlation.
6. Consistency Quality Control Pathway Index (CQCp): This metric is similar to CQCg but uses enrichment pathways.

A PCA plot was generated, where the direction of each QC measure was juxtaposed on top of the two-dimensional PC subspace using arrows. Datasets in the negative region of the arrows were classed as outlier studies (Kang *et al.*, 2012).

1.8.2 Microarray gene expression analysis

Genes are functional units of genetic material, which have been shown to interact with other genes for biological functionality. Therefore, this thesis focused on the identification of DEG's and modules of highly co-expressed genes in AD, which collectively may provide information on biological functions disrupted in the disease. The following is a detailed description of the analysis used to identify these genes and analyses for biological interpretation. Further details of the application used to perform the analysis is provided in individual chapters.

1.8.2.1 Differential expression analysis

Differential expression analysis is a method to identify significantly perturbed genes between two or more different biological conditions. The analysis was performed using the Linear Models for Microarray Data (Limma) R package, which fits a linear model to each gene in the data (Smyth, 2005). During the analysis, only sex and diagnosis were used as covariates, with the assumption SVA and COMBAT had removed any latent variation.

1.8.2.2 A meta-analysis of gene expression datasets

Replication of DEG's across independent transcriptomic microarray studies remains poor (Zhang *et al.*, 2008), with the primary cause thought to be due to a small number of samples leading to low statistical power (Chang *et al.*, 2013). Combining information across studies can be an intuitive way to increase sensitivity and can be accomplished through a p-combining meta-analysis. Many meta-analysis methods exist for microarray data, with a review on 12 different methods concluding Adaptively Weighted (AW) as one of the best methods to identify DEG with non-zero effect sizes in one or more microarray studies (Chang *et al.*, 2013). AW has additional advantages by indicating which datasets are contributing to the meta-analysis, is biased towards studies with concordant significant effects (Li and Tseng, 2011), and has been demonstrated to be robust with identified DEGs showing relevant biological associations (Chang *et al.*, 2013).

1.8.2.3 Weighted Gene Correlation Network Analysis

Correlation networks can be used to identify clusters of highly correlated genes, which collectively may represent or be involved in biological functions. Weighted Gene Correlation Network Analysis (WGCNA) is a method to identify such clusters and was used in this thesis. First, a "soft thresholding power" based on the criterion of approximate scale-free topology was calculated from the data and was then used to calculate the "signed" gene adjacency values (Langfelder and Horvath, 2008). Then, hierarchical clustering of the adjacency values was performed to create a dendrogram, where densely interconnected branches represent highly co-expressed genes. Individual branches were assigned to so-called "modules" (Langfelder and Horvath, 2008).

1.8.2.4 Protein-Protein Interaction network analysis

Proteins are essential to cellular and molecular functions and rarely act alone. Therefore, the physical interactions between proteins can provide insight into cellular and molecular mechanisms in AD (Safari-Alighiarloo *et al.*, 2014). Protein-Protein Interactions (PPI) networks were independently created on DEG's lists and on co-expressed gene modules using a web-based application called NetworkAnalyst (Xia, Benner and Hancock, 2014). The application annotates the input gene list to their relevant proteins, searches binary interactions form curated PPI databases and creates a network of these interactions. A topology analysis was then performed on the whole network to identify important nodes (hubs) which represent highly interconnected proteins that may play an important role in cellular signalling (Xia, Benner and Hancock, 2014).

1.8.2.5 Gene Set Enrichment Analysis

GSEA is an alternative method to interpret gene expression data based on functional annotation of DEG's or modules of co-expressed genes. GSEA can provide an insight into biological processes or

molecular functions that may be altered in a disease, which has been shown to be more versatile and reproducible across similar studies than DEG's alone (Subramanian *et al.*, 2005). A number of biological pathway databases exist, each consisting of valuable, often manually curated, and validated functional interactions between genes. ConsensusPathDB is a web-based application that comprehensively queries up to 30 publicly available repositories using a hypergeometric test, compiles a set of results and corrects for multiple testing (Kamburov *et al.*, 2009). ConsensusPathDB was used throughout this thesis for biological pathway and Gene Ontology (GO) interpretations.

1.8.3 Publicly available transcriptomic data

A vast quantity of microarray-based gene expression data is increasingly becoming publicly available through repositories such as ArrayExpress (https://www.ebi.ac.uk/arrayexpress/), Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) and an AD focused Accelerating Medicines Partnership-Alzheimer's Disease (AMP-AD, https://www.nia.nih.gov/research/amp-ad). These databases have stringent procedures to store only well-annotated data in a structured manner and generally store data from all microarray expression platforms. These three repositories were the largest known resources for publicly available microarray gene expression cohorts relevant to this thesis. These repositories were queried for suitable datasets and are discussed within individual analysis chapters.

1.8.4 Machine Learning

Machine learning (ML) is the use of algorithms and statistical models by computer systems to learn, improve from experience and predict an outcome automatically. There are two primary ML techniques; supervised and unsupervised. Supervised ML is useful for regression and classification on labelled data, while unsupervised ML is generally applied to cluster unlabelled data. This thesis used a supervised approach to build a classification model for biomarker discovery.

For classification purposes, ML is developed on one dataset and ideally validated in an independent dataset, which has not been seen by the ML algorithm. The dataset used for the classification model development is referred to as the "training set", while the validation dataset is referred to as the "testing set".

1.8.4.1 XGBoost overview

Decision trees are a type of ML algorithms that are best suited to solve classification problems, and as the name implies, are trees of decisions. Tree boosting is a series of simple decision trees, where each successive tree is developed from the residuals of the preceding tree. One of the most widely used tree boosting algorithms that have shown to give the most accurate results on many classification benchmarks is Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016). In addition, XGBoost supports weighted samples for unbalanced datasets, uses parallel and distributed computing for fast

model exploration of large datasets, and allows many hyperparameters to be tuned to reduce overfitting. Overall, this made XGBoost an ideal algorithm to use in this thesis.

XGBoost converts a series of weak learners, which are slightly better than random, into a collectively strong learner. This is achieved by using information from the prior tree to improve prediction in the following predictions. The overall algorithm idea is illustrated below in Figure 1.4:

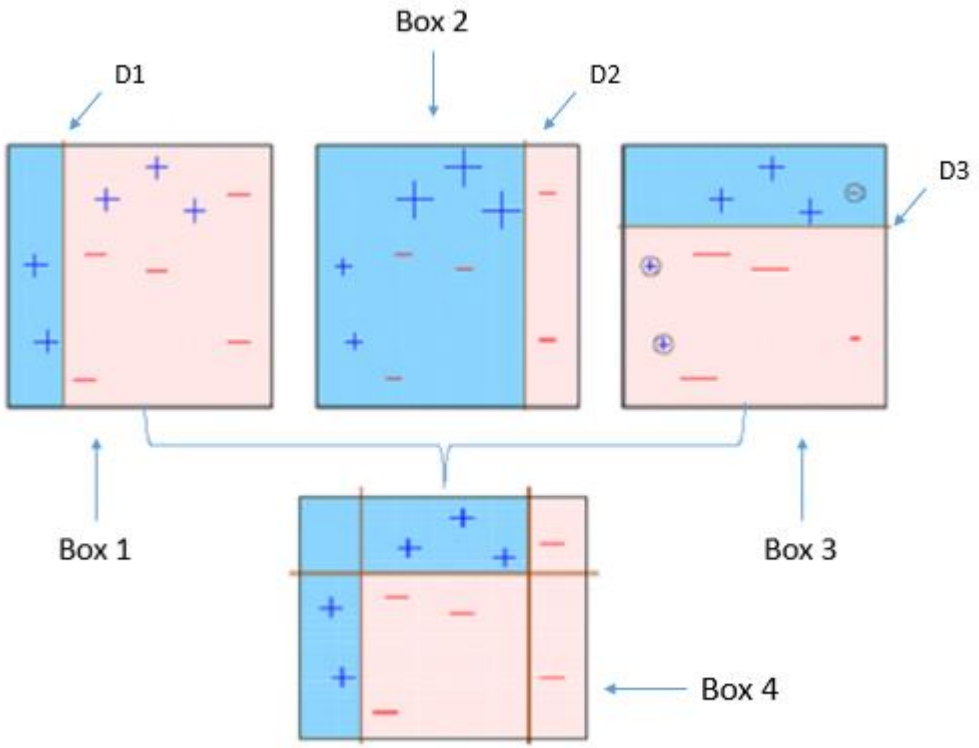


Figure 1.4: The Basic idea behind the XGBoost algorithm. Four trees are represented by 4 boxes, which are trying to classify + and – classes to hypothetical biological samples. The blue shaded regions of each box are samples classified as the + class, while the red shaded region is assigned a – class. The first row of boxes misclassifies samples and can be individually regarded as weak learners; however, collectively (Box 4) they correctly classify all samples. Reproduced from (Manish Pathak, 2018).

Four trees represented by 4 boxes are shown above in Figure 1.4, which are trying to classify + and – classes to hypothetical biological samples. The tree boosting algorithm follows the basic idea to develop a strong classifier from weak learners:

Box 1: The first classifier creates a vertical line at D1 to classify samples to the left of the line as + and samples to the right as -. This classifier misclassifies three + samples and can be considered a weak classifier.

Box 2: The next classifier attempts to correct the previous classifiers (Box1) mistakes by assigning more weights to the three + misclassified points (bigger size of +). The new classifier creates a vertical line at

D2 to classify samples to the left of the line as + and samples to the right as -. This classifier misclassifies three - samples and can be considered a weak classifier.

Box 3: The next classifier continues to improve upon the previous by assigning more weight to the three – samples that were misclassified and creates a horizontal line at D3. This classifier misclassifies three samples (circled) and can be considered a weak classifier.

Box 4: Is a weighted combination of all the weak classifiers, which performs considerably better than individually, and be considered a strong classifier.

This basic concept makes XGBoost a powerful tool for developing classification models. However, due to the simplicity of the concept, the ML algorithm can easily create an overfitted model, which is when the algorithm learns random error in the data rather than relationships between variables. To control for overfitting, XGBoost incorporates cross-validation and allows for many hyperparameters to be finely tuned to reduce the chance of overfitting.

1.8.4.2 K-fold cross-validation

Cross-validation (CV) is an approach that can be used to estimate the performance of a classification model. The data is divided into a number (n) of 'folds', where n-1 folds are used to develop the classification model, and the remaining fold used for validation purposes. This process is repeated n times where each fold is used as a validation set only once. The average validation performance across all folds is used to determine classification model performance and has been shown to reduce the chance of developing an overfitted model (Tušar *et al.*, 2017).

1.8.4.3 Model tuning and Logloss

The XGBoost algorithm allows many parameters to be adjusted to improve classification model performance, whilst reducing the chances of overfitting. This process is generally referred to as "tuning" and was performed in a sequential manner whilst performing CV. The CV process was deployed using the standard XGBoost xgb.cv function to identify the optimum hyperparameters, which uses the default k-fold CV rather than a nested-cross validation. The optimum parameter configurations were selected based on the logloss error obtained through the CV. Logloss error is a performance metric that quantifies the accuracy of a classification model by penalising errors in prediction whilst taking into consideration how uncertain the prediction is. Hyperparameters were adjusted until the lowest logloss error was obtained.

Chapter 2: Transcriptomic analysis of probable asymptomatic and symptomatic Alzheimer brains

2.1 Background

The increase in life expectancy has profoundly increased the ageing population, which, unfortunately, is also accompanied by a rise in age-related disorders including Alzheimer's disease (AD) (Prince *et al.*, 2016). Alzheimer's disease is a neurodegenerative disorder characterised by progressive accumulation of extracellular amyloid-β (Aβ) protein and intracellular hyperphosphorylated tau filaments in the brain, which form insoluble plaques and tangles respectively. These protein aggregates affect neuronal activity, which can lead to progressive loss of neurons associated with deterioration in cognition and development of neuropsychiatric symptoms.

Through longitudinal studies involving autopsy, it has become evident that clinical signs of cognitive impairment are apparent after substantial years of neurodegeneration, which occurs decades after neuropathological changes (Caselli and Reiman, 2012). As the disease is progressively slow, and as everyone is expected to experience cognitive change during normal ageing, differentiating AD symptoms from normal ageing at an early stage of the disease can be difficult. Up to 20-30% of the ageing population with intact cognition have amyloid deposition, with these individuals at higher risk of progressing to AD than those without amyloid (Yvette *et al.*, 2010). These individuals are often referred to as asymptomatic AD (AsymAD) (Driscoll and Troncoso, 2011) and have been shown to be distinguishable from normal ageing based on neuropathology, brain imaging and cerebrospinal fluid biomarkers (Caselli and Reiman, 2012). While some of these individuals progress to developing symptoms related to cognition, which deviate from normal Mild Cognitive Impairment (MCI), and then to AD, not all do. They are therefore a heterogeneous group, representing those with prodromal AD and those impervious to AD despite having the pathological hallmarks.

Measuring genome-wide expression of transcripts as markers of gene activity has revealed that cognitive decline is accompanied by changes in brain gene expression from normal ageing through to MCI and AD. Studies have suggested that some changes in the pattern of gene expression in normal ageing such as synaptic function and energy metabolism (Saura, Parra-Damas and Enriquez-Barreto, 2015), are extensively altered in MCI (Nicole C. Berchtold, Ph.D.1, Marwan N. Sabbagh, M.D.2, Thomas G. Beach, M.D. *et al.*, 2015) and AD (Blalock *et al.*, 2004, 2011; Miller, Oldham and Geschwind, 2008; Berchtold *et al.*, 2013; Miller *et al.*, 2013) Additional work has also suggested a number of other biological pathways are more specifically altered in AD, including inflammation (Colangelo *et al.*, 2002; Blalock *et al.*, 2004; Miller, Oldham and Geschwind, 2008), protein misfolding (Colangelo *et al.*, 2002;

Blalock *et al.*, 2004), transcription factors (Colangelo *et al.*, 2002; Blalock *et al.*, 2004), cell proliferation (Colangelo *et al.*, 2002; Blalock *et al.*, 2004), immune response (Lambert *et al.*, 2010; Li *et al.*, 2012; Sekar *et al.*, 2015; Chen *et al.*, 2016), protein transcription/translation regulation (Li *et al.*, 2012, 2015; Liu *et al.*, 2013; Godoy *et al.*, 2014; Sekar *et al.*, 2015; Puthiyedth *et al.*, 2016), calcium signalling (Blalock *et al.*, 2011; Ramanan *et al.*, 2013; Sekar *et al.*, 2015), MAPK signalling (Miller *et al.*, 2013; Puthiyedth *et al.*, 2016), and various metabolism pathways (Ishii *et al.*, 1997; Liang *et al.*, 2008; Oshiro, Morioka and Kikuchi, 2011; Li *et al.*, 2012; Paolo and Kim, 2012; Godoy *et al.*, 2014; Puthiyedth *et al.*, 2016) which reflect the extent and type of pathology and disruption to cell activity as disease progresses. It is unknown how early the different types of changes occur in the brain, such as in the pre-symptomatic phase or specifically in AsymAD subjects who already have the pathological hallmarks of AD such as amyloid and neurofibrillary tangles (NFTs). Understanding the fundamental changes in this AsymAD group may shed light on specific biological mechanisms that may be involved in early pathological hallmarks of AD, providing new therapeutic targets for early intervention.

This study investigated transcriptomic changes in the human brain of healthy ageing, AsymAD and AD subjects, which have been classified based on the clinical assessment before death and AD neuropathology at autopsy. Typical transcriptomic analysis coupled with a systems-biology approach was used to identify disturbances in the underlying biological mechanisms across the entorhinal cortex, temporal cortex, frontal cortex and cerebellum brain regions. In addition, access of gene-level results to the broader research community is provided through a publicly available R SHINY web-application (https://phidatalab-shiny.rosalind.kcl.ac.uk/ADbrainDE), allowing researchers to quickly query the expression of specific genes through the progression of AD and across multiple brain regions.

2.2 Materials and Methods

2.2.1 Medical Research Council London Neurodegenerative Diseases Brain Bank

A total of 112 brains were obtained from the Medical Research Council (MRC) London Neurodegenerative Diseases Brain Bank (from now on referred to as MRC-LBB) hosted at the Institute of Psychiatry, Psychology and Neuroscience, KCL. All cases were collected under informed consent, and the bank operates under a licence from the Human Tissue Authority, and ethical approval as a research tissue bank (08/MRE09/38+5). Neuropathological evaluation for neurodegenerative diseases was performed in accordance with standard criteria.

2.2.2 MRC-LBB sample selection

BRAAK staging is a measure of the spread of hallmark AD pathology across the brain and is part of the neuropathological assessment. In general, BRAAK stages I-II, III-IV and V-VI have been suggested to represent prodromal, early-moderate AD, and moderate-late AD respectively. Twenty-seven control

cases were used - classified as showing no clinical sign of any form of dementia and no neuropathological evidence of neurodegeneration. Thirty-three AsymAD cases were also analysed - defined as clinically dementia-free at the time of death, but neuropathological assessment at autopsy showed hallmark AD pathology. Finally, fifty-two AD cases, which had both a clinical diagnosis of AD at death and confirmation of this diagnosis through neuropathological evaluation at autopsy, were selected.

2.2.3 MRC-LBB brain region selection and RNA extraction

Frozen tissues (0.5-1cm3) from the following brain regions from each case were macrodissected into RNAlater RNA Stabilization Reagent (Qiagen): 1) Frontal Cortex (FC), 2) Temporal Cortex (TC), 3) Entorhinal Cortex (EC) and 4) Cerebellum (CB). Hallmark AD pathology was confirmed in the entorhinal cortex, temporal cortex and frontal cortex but absent from the cerebellum of AsymAD and AD subjects. RNA extraction was performed within 24 hours of dissection. Total RNA was extracted using RNeasy Lipid Tissue Mini Kit (Qiagen,74804) following the manufacturer's protocol. Genomic DNA was removed using gDNA Eliminator Spin Columns (Qiagen). The RNA quality was evaluated with an Agilent 2100 bioanalyzer (Agilent Technologies, Inc., Palo Alto, CA).

2.2.4 MRC-LBB Illumina beadArray expression profiling

Total RNA (25ng) was prepared for array expression profiling using the Ovation Pico WTA System (NuGEN Technologies, Inc., San Carlos, CA), as described by the manufacturer's protocol. The Nugen system is optimised for the amplification of degraded RNA, where amplification is initiated at the 3' end as well as randomly throughout the whole transcriptome. The samples were processed at the NIHR Biomedical Research Centre for Mental Health (BRC-MH), Genomics & Biomarker Core Facility at the Social, Genetic & Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London (https://www.kcl.ac.uk/ioppn/depts/sgdp-centre/research/The-IoPPN-Genomics--Biomarker-Core-Facility.aspx) in accordance with the manufacturer's protocol using the Illumina HT-12_V4 beadchips (Illumina, USA).

2.2.5 Microarray expression data processing

Raw gene expression data was exported from Illumina's GenomeStudio (version 2011.1) into RStudio (version 0.99.467) for data processing. Using R (version 3.2.2), raw data was Maximum Likelihood Estimation (MLE) background corrected using R package "MBCB" (version 1.18.0) (Allen, Chen and Xie, 2010), log2 transformed, and underwent Robust Spline Normalisation (RSN) using R package "lumi" (version 2.16.0) (Du, Kibbe and Lin, 2008).

A series of quality control steps were carried out before data analysis. Duplicate samples were removed based on lowest RIN score. Sex was predicted for each sample using the R package "massiR" (version

1.0.1) (Buckberry *et al.*, 2014), with any discrepancies in predicted and clinically recorded sex from the same individual across all tissues removed from further analysis. For each sample, probesets "not reliably detected" or "unexpressed" were removed to eliminate noise (Lazar *et al.*, 2013) and increase power (Hackstadt and Hess, 2009). If the expression of a probe was below the 90th percentile of the log2 expression scale in over 80% of samples across all groups (based on disease status, brain region and sex), the probe was deemed "unexpressed" and was removed from further analysis.

Batch effects were then explored using Principal Component Analysis (PCA) and Surrogate Variable Analysis (SVA) using the R package "sva" (version 3.10.0) (Leek *et al.*, 2012). Sex and diagnosis information was used as covariates in sva when correcting for unknown batch effects. To ensure homogeneity among the biological groups, outlying samples per tissue and disease group were iteratively identified and removed following the fundamental network concepts described in (Oldham, Langfelder and Horvath, 2012). Finally, Illumina-specific probe ID's were converted to the universal Entrez Gene ID using the R package "illuminaHumanv4.db" (version 1.22.1).

2.2.6 Differential Expression and Gene Set Enrichment Analysis

Differential Expression (DE) analysis was performed using the R package "limma" (version 3.20.9) (Smyth, 2005). As unwanted batch effects in the data were theoretically addressed using sva, only used sex in the DE model was used as a covariate. A gene was regarded as significantly differentially expressed if the false discovery rate (FDR) adjusted p-value ≤ 0.05.

Gene set enrichment analysis (GSEA) was performed using an Over-Representation Analysis (ORA) implemented through the ConsensusPathDB (http://cpdb.molgen.mpg.de) web-based platform (version 32) (Kamburov *et al.*, 2009) in October 2017. ConsensusPathDB incorporates numerous well-known biological pathway databases including BioCarta, KEGG, Reactome and Wikipathways. It performs a hypergeometric test while combining a background gene list, compiles results from each database and corrects for multiple testing using FDR (Kamburov *et al.*, 2009). During GSEA analysis, a minimum overlap of the query signature and database was set as 2.

2.2.7 Weighted Gene Co-expression Network Analysis

Weighted gene co-expression analysis (WGCNA) was performed using R package "WGCNA" (version 1.51) to identify clusters (modules) of highly correlated genes, with the underlying hypothesis that such modules could possess a common function. The WGCNA analysis was performed as described in (Langfelder and Horvath, 2008). In brief, a co-expression network based on "signed" adjacency was independently created for all three phenotypes (control, AsymAD and AD group), topological overlap calculated, and hierarchical clustering used to group genes into modules. The control group module was assigned default colours based on module size, and the AsymAD and AD module colours

determined based on the control module gene overlap. Module cross-tabulations were generated across the three phenotypes and Fisher's exact test used to test for enrichment between modules-gene assignments between the control, AsymAD and AD groups. To aid in identifying significant changes in the co-expression network within the same modules in the three phenotypes, additional statistics known as "Module preservation Zsummary" and "median rank" were calculated as described in (Langfelder *et al.*, 2011).

2.2.8 Protein-Protein Interaction network analysis

Protein-protein interaction (PPI) networks were generated by uploading gene lists (referred to as seeds in network analysis) to NetworkAnalyst's (http://www.networkanalyst.ca/faces/home.xhtml) web-based platform in December 2017. The "zero-order network" option was incorporated to allow only seed proteins directly interacting with each other, preventing the well-known "hairball effect" and allowing for better visualisation and interpretation (Xia, Benner and Hancock, 2014). Sub-modules with a p-value ≤ 0.05 based on the "InfoMap" algorithm (Zaki and Mora, 2015) were deemed significant "hubs" and the gene(s) with the most connections within this network as the "key hub gene(s)".

2.2.9 Study design

Differential and co-expression analysis was performed between the three disease groups and for each of the four brain regions. First, the control and AsymAD groups were compared, and from this point onwards is referred to as the "**Early AD**" analysis. Second, the AsymAD and AD groups were compared, and from this point onwards is referred to as the "**Late AD**" analysis. Finally, the control and AD groups were compared, and from this point onwards is referred to as the "**Standard AD**" analysis. An overview of the study design and analyses is shown in Figure 2.1.

2.2.10 Data availability

The microarray data have been deposited in NCBI's GEO database under the accession number GSE118553. Additionally, a shiny application was written in R using the "shiny" framework (version 0.14) to allow quick visualisation of specific gene expression in the control, AsymAD and AD subjects, and across the EC, TC, FC and CB brain regions. The application also displays DE results of each gene and can be accessed at https://phidatalab-shiny.rosalind.kcl.ac.uk/ADbrainDE. All data analysis scripts used in this study are available at https://doi.org/10.5281/zenodo.1400644

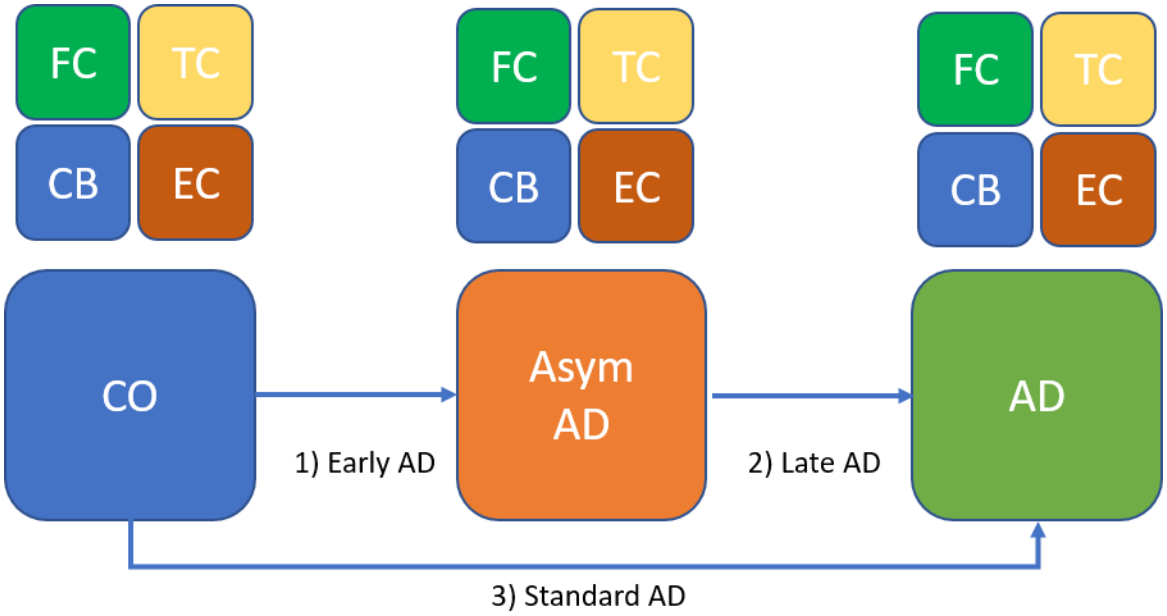


Figure 2.1: Overview of Study Design. Four brain regions; frontal cortex (FC), temporal cortex (TC), entorhinal cortex (EC) and cerebellum (CB) from the three subject groups; control (CO), Asymptomatic AD (AsymAD) and Alzheimer's disease (AD) were expression profiled. The typical comparison between the CO and AD group is referred to as the "Standard AD" analysis, the comparison between the CO and AsymAD group is referred to as the "Early AD" analysis and the comparison between the AsymAD and AD group is referred to as the "Late AD" analysis.

2.3 Results

2.3.1 Data processing

Of the 401 tissue samples assessed (extracted from the 112 brains) 48 samples were removed due to duplication, 4 samples due to outlier detection analysis and 2 samples due to sex discrepancies between recorded and actual sex, leaving 347 tissue samples from 111 brains for DE and co-expression analysis. As a result of samples not being microarray profiled due to sampling quality, and samples being removed during the Quality Control (QC) process, not all subjects had tissue samples extracted from all four brain regions. The demographics for datasets by brain region and sample group is provided in Table 2.1.

After further QC and annotation to determine Entrez gene identifiers, the final data represented 3518 "reliably detected" genes across all samples. Chi-squared tests revealed no significant difference in the proportion of males to females across the three disease groups or brain regions. Mann-Whitney U test revealed no significant difference between post-mortem (PM) delay or disease duration across analyses; however, age was significantly ($p \leq 0.01$) lower in the control groups when compared to the AsymAD and AD group in each tissue (see Supplementary Table 2.1).

2.3.2 Summary of differentially expressed genes across disease groups and tissues

A summary of DEG's identified in each brain tissue and analyses are illustrated in Figure 2.2, and a full list of DEG's can be accessed: https://phidatalab-shiny.rosalind.kcl.ac.uk/ADbrainDE. The general trend of DEG's in subjects with AD ("Late AD" and "Standard AD" analysis) decreases across brain regions in the order of EC (n=1904 and n=1690 respectively) > TC (n=1546 and n=1517 respectively) > FC (n=52 and n=299 respectively) > CB (n=13 and n=176 respectively). This expression pattern corresponds to the route AD pathology is seen to spread through the brain. By contrast, the pattern differs in the AsymAD group ("Early AD" analysis), where most DEGs are detected in the FC (n=398) followed by the TC (n=253), EC (n=19) and CB (n=1), suggesting initial molecular changes may begin in the FC brain region prior to AD symptoms.

Table 2.1: Summary of MRC-LBB sample characteristics

Brain Region	Phenotype	No. Samples	Sex (M/F)	Age (± SD)	BRAAK (± SD)	PM Delay (h)	Disease Duration (yrs)
Entorhinal Cortex	Control	16	9/7	71.9 (15.6)	0	33.8 (17.8)	0
	AsymAD	28	8/20	85.4 (9.5)	2.2 (± 1.2)	52.5 (15.9)	0
	AD	34	13/21	83.9 (9.7)	4.9 (± 1)	39.5 (21.2)	11.8 (5.2)
Temporal Cortex	Control	24	14/10	71.5 (16.9)	0	37.2 (19.8)	0
	AsymAD	28	9/19	86.3 (8.6)	2.5 (± 1.1)	54.2 (16.6)	0
	AD	45	20/25	82.7 (9.8)	4.9 (± 0.9)	40.4 (21.4)	9.7 (5.4)
Frontal Cortex	Control	21	12/9	69.8 (15.4)	0	40.4 (24.6)	0
	AsymAD	32	10/22	86 (8.9)	2.3 (± 1.2)	54.1 (16.2)	0
	AD	38	13/25	82.5 (4.7)	4.9 (± 1)	39.4 (20.5)	10.5 (5.7)
Cerebellum	Control	18	10/8	69.4 (16)	0	37.9 (20.7)	0
	AsymAD	27	8/19	86.3 (9.2)	2.4 (± 1.2)	56 (16.5)	0
	AD	36	17/19	82.6 (10.6)	5.1 (± 0.3)	40.2 (22.3)	9.4 (5.6)

The table provides a summary of sample characteristics used in this study. From the Initial 401 samples expression profiled, 48 samples were removed due to duplication, 2 samples removed due to sex discrepancies and 4 samples removed due to being identified as outliers. The total number of samples available after quality control was 347. BRAAK staging is a measure of the spread of hallmark AD pathology across the brain and does not reflect pathology within a distinct brain region. In general, BRAAK stages I-II, III-IV and V-VI have been suggested to represent prodromal, early-moderate AD, and moderate-late AD respectively. BRAAK scores deviate between brain regions as not all four brain regions were available from all donors. Hallmark AD pathology was confirmed in the entorhinal cortex, temporal cortex and frontal cortex but absent from the cerebellum of AsymAD and AD subjects. The values provided in Age, BRAAK and PM Delay represent the mean ± standard deviation. Abbreviation: M/F: the ratio of male and female samples, PM: Post-Mortem, h: Hours, yrs: Years, SD: Standard deviation.

2.3.2.1 AD tau pathology marker suggests AsymAD subjects are an Intermediate state between normal ageing and AD.

A previous study identified eight genes highly correlated with AD tau pathology (Miyashita *et al.*, 2014), of which two genes (**RELN, TRIL**) are present in the data. DE analysis results indicate the **TRIL** gene

expression gradually increases through the control, AsymAD and then the AD group. In addition, the expression increase is only observed in brain regions known to be affected by tau pathology (EC, TC and FC), and the extent of expression change within these affected brain regions shadows the route of disease manifestation through the brain (Figure 2.3a). The EC exhibits the most significant increase of **TRIL** expression (logFC=0.99, FDR adjusted p-value=2.77e-8), followed by the TC (logFC=0.48, FDR adjusted p-value=1.41e-3) and then FC brain region (logFC=0.44, FDR adjusted p-value=2.21e-2). This expression pattern further suggests the **TRIL** gene is a reliable brain marker for tau pathology, and the AsymAD samples are a good representation of early-intermediate state between normal ageing and AD.

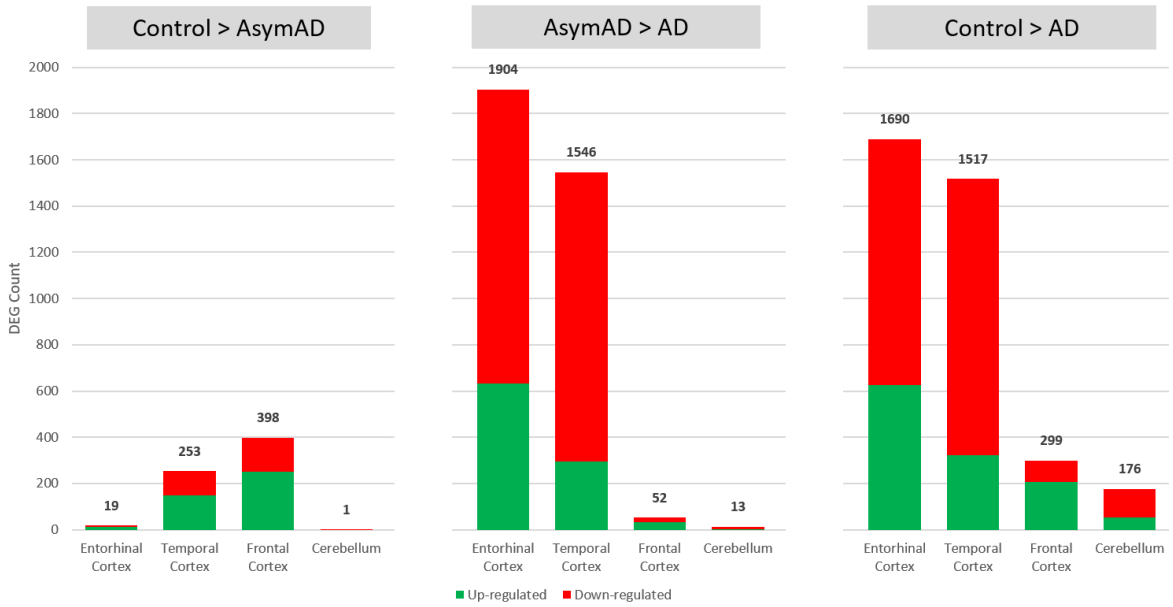


Figure 2.2: Distribution of Significant DEG (FDR adjusted p-value ≤ 0.05) across brain regions and analyses. "Control>AsymAD" summarises the number of DEGs between the control and AsymAD group. "AsymAD>AD" summarises the number of DEGs between the AsymAD and AD group. "Control>AD" summarises the number of DEGs between the control and AD group. The proportion of up-regulated genes is represented in green while the down-regulated genes are represented in red. The total number of significantly differentially expressed genes in each brain region and analysis is provided on top of each bar. More genes are observed to be generally perturbed when comparing the AD group to the AsymAD or healthy ageing group, with the general pattern of more genes perturbed in the entorhinal cortex, followed by the temporal cortex, frontal cortex and then the cerebellum, a pattern generally representing the spread of hallmark AD pathology. In contrast, comparing the AsymAD group to the healthy ageing group reveals more genes are perturbed in the frontal cortex, followed by the temporal cortex, entorhinal cortex and then the cerebellum, suggesting initial molecular changes in AD may begin in the frontal cortex before the manifestation of clinical AD symptoms.

2.3.2.2 The most significant differentially expressed genes per analysis

The most DEG's from each analysis is 1) **MOSPD3** (downregulated in the TC brain region in "Early AD", FDR adjusted p-value = 1.18e-10, Figure 2.3b), 2) **NPC2** (upregulated in the EC brain region in the "Late

AD" analysis, FDR adjusted p-value = 2.39e-20, available to view in the SHINY web-app) and 3) NOTCH2NL (upregulated in the EC brain region in the "Standard AD" analysis, FDR adjusted p-value = 1.29e-15, available to view in the SHINY web-app).

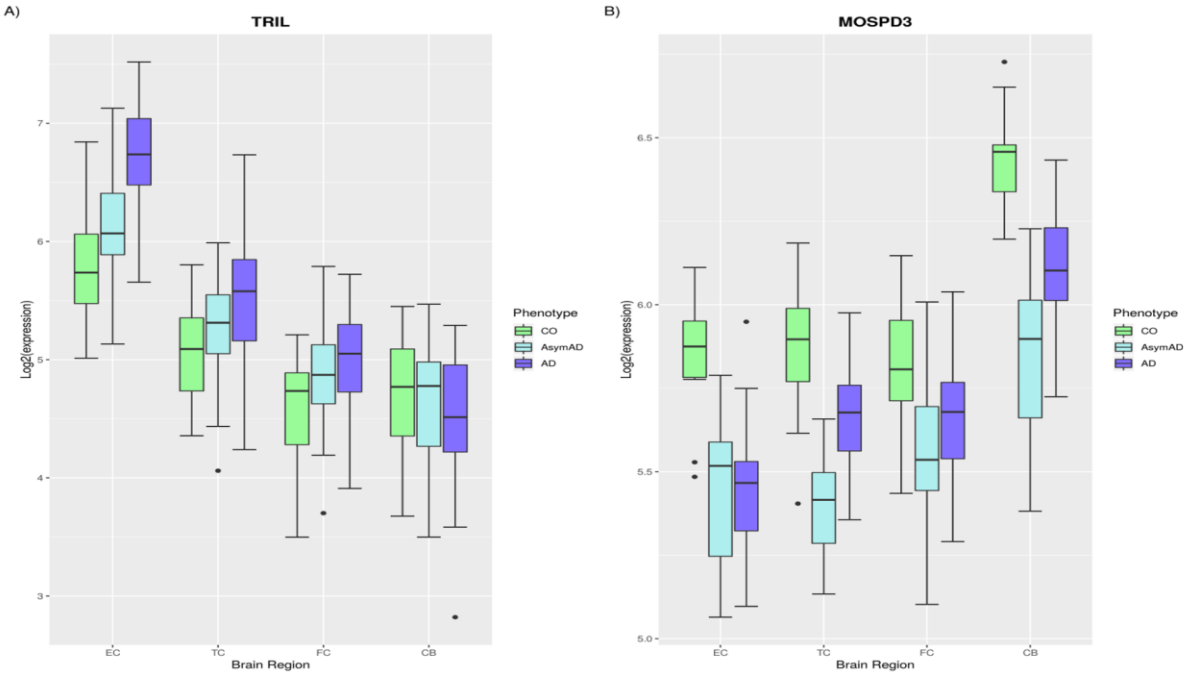


Figure 2.3: Expression boxplots of the TRIL and MOSPD3 genes. A) Previous research identified the TRIL gene as a marker for Tau pathology. The TRIL gene is significantly up-regulated (before multiple correction) from control to AsymAD (EC: logFC= 0.39 & p-value=0.01, TC: logFC=0.24 & p-value=0.04, FC: logFC=0.24 & p-value=0.04) and then further to AD (EC: logFC= 0.6 & p-value=6.57e-6, TC: logFC=0.29 & p-value=4.19e-3, FC: logFC=0.18 & p-value=0.05), but not in the cerebellum (control to AsymAD: logFC= -0.01 & p-value=1, AsymAD to AD: logFC=-0.19 & p-value=0.2), a region spared by hallmark AD pathology. The expression pattern of the TRIL gene further supports the assignment of AsymAD samples, which were based on clinical records and neuropathological assessment, as an early intermediate state between healthy ageing and AD. B) The MOSPD3 gene is the most Significant DE gene in the Early AD analysis and is consistently down-regulated in all brain regions of the AsymAD group when compared to controls (EC: logFC=-0.38 & adjusted p-value=5.6e-4, TC: logFC=-0.5 & adjusted p-value=1.18e-10, FC: logFC= -0.27 & adjusted p-value=6.91e-4, CB: logFC=-0.59 & adjusted p-value= 1.51e-6). As all brain regions are affected in AD, albeit not to the same degree, the MOSPD3 gene may represent an early brain biomarker for cell dysfunction in AD.

2.3.2.3 Common differentially expressed genes across all brain regions

The overlap of DE genes across brain regions is shown in Figure 2.4. MOSPD3 is the only gene significantly differentially expressed across all four brain regions in the "Early AD" analysis. No gene was significantly differentially expressed in the "Late AD" analysis across all four brain regions; however, six genes (NPC2, DUSP1, GPM6B, SLC38A2, ANKEF1, MOSPD3) were identified in "Standard AD" analysis. Three of these genes (DUSP1, SLC38A2 and MOSPD3) are consistently expressed in the same direction across all four brain regions. DUSP1 and SLC38A2 gene expression are upregulated during disease progression (Control to AsymAD to AD). MOSPD3, however, is downregulated in the disease in both the "Early AD" and "Standard AD" analyses, with no significant difference between the AsymAD and AD

subjects. The remaining three genes (**NPC2, GPM6B, ANKEF1**) are DE in the same direction across all brain regions but reversed in the CB; a brain region suggested to be spared by hallmark AD pathology.

2.3.2.4 Differentially expressed genes in brain regions with hallmark AD pathology

The EC, TL and FC are all affected by hallmark AD pathology (amyloid and NFT's), while the CB is known to be partially spared. Gene's DE in the EC, TC and FC brain regions and not the CB, may identify hallmark AD pathology specific genes.

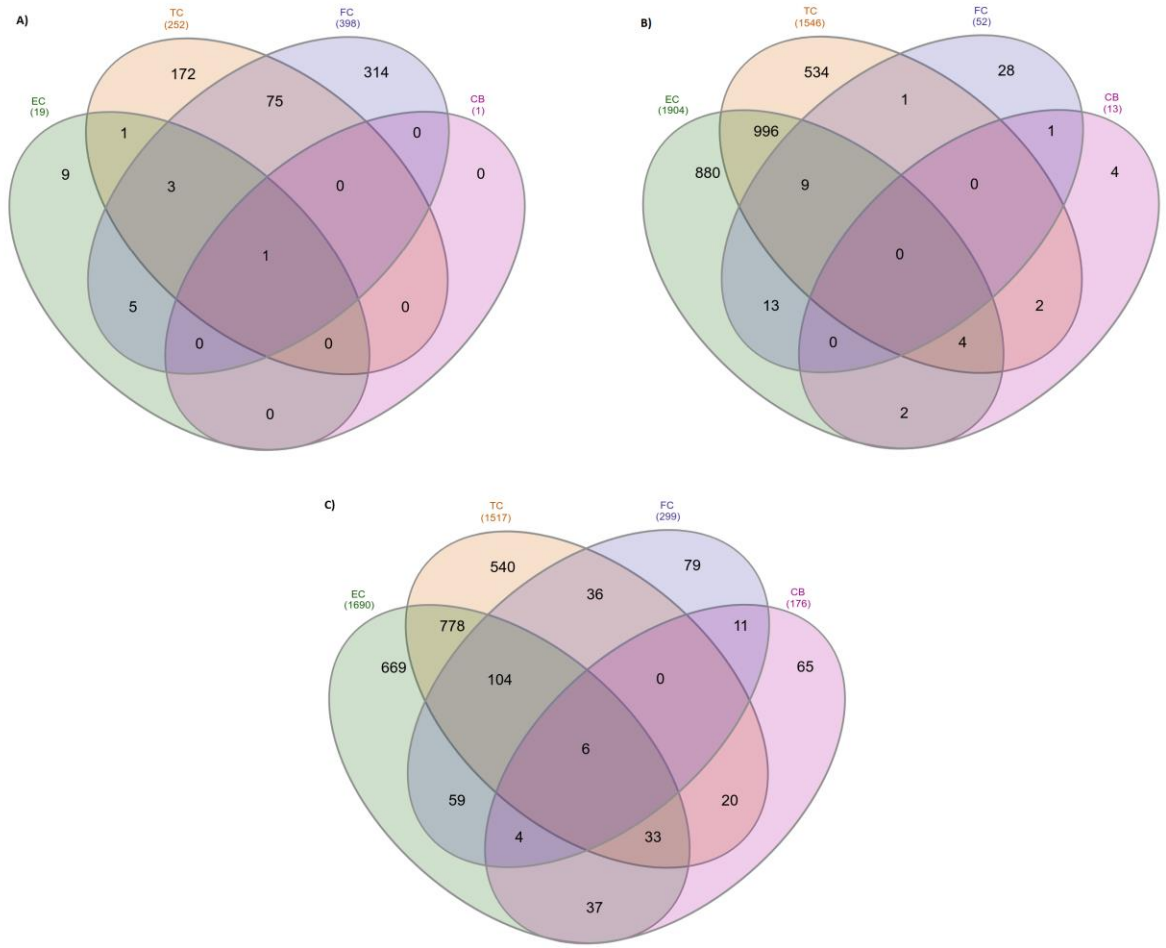


Figure 2.4: Overlap of significant DEG across brain regions in A)" Early AD" analysis, B) "Late AD" analysis and C) "Standard AD" analysis. All brain regions in this study are affected in AD, specifically by atrophy and neuronal loss, while only three brain regions in this study (EC, TC and FC) are affected by the additional accumulation of hallmark AD pathology (Aβ and NFT). Genes perturbed across all brain regions may be markers of cell dysfunction in AD, while genes consistently perturbed in the EC, TC and FC but not in the CB may be associated with AD pathology. ***MOSPD3*** *gene is the only gene DE in all brain regions of the "Early AD" analysis. No gene is DE across all brain regions in the "Late AD" analysis. Three (****ALDH2, FBLN2****,* ***METTL7A****) and nine genes (****FLCN, ASPHD1, ARL5A, GPR162, HBA2, PCID2, NDRG2, BEND3, RAP1Gap****) are consistently DE across all brain regions affected by hallmark AD pathology in the "Early AD" and "Late AD" analyses respectively.*

Three (ALDH2, FBLN2 and METTL7A) and nine (FLCN, ASPHD1, ARL5A, GPR162, HBA2, PCID2, NDRG2, BEND3, RAP1Gap) genes were significantly differentially expressed across the EC, TC and FC brain regions and not the CB brain region in the "Early AD" and "Late AD" analysis respectively.

2.3.2.5 Gene Set Enrichment Analysis of differentially expressed genes

To understand the functional implications of DEG's, GSEA was performed using the significant DEG list from all three analyses ("Early AD", "Late AD" and "Standard AD") and across all four brain regions. No biological pathway is significantly enriched across all four brain regions in the "Early AD", "Late AD" or "Standard AD" analysis. However, when excluding the brain region often referred to spared by hallmark AD pathology (CB), the "glutamate glutamine metabolism" and "gluconeogenesis and glycolysis" pathways are the only pathways significantly enriched in the "Early AD" and "Late AD" analysis respectively. For the "Standard AD" analysis, excluding the CB brain region additionally identified "mRNA processing", "synaptic vesicle pathway" and "TNF-alpha" pathways as significantly enriched in the remaining three brain regions.

2.3.3 Summary of Weighted Co-Expression Network Analysis

Weighted gene co-expression analysis was performed on the FC and EC brain regions. This study focused on these two brain regions as differential expression analysis identified an increased number of significant DEG's in the FC brain region prior to AD symptoms and the EC is widely regarded as one of the first areas of the brain to be affected in AD. Network preservation and cross-tabulation statistics were calculated to identify co-expression networks that may be preserved or disrupted between the Control, AsymAD and AD subjects. Figure 2.5 illustrates the WGCNA module assignments and module preservation statistics, and Figure 2.6 shows the cross-tabulation statistics across phenotypes.

Co-expression analysis in the FC brain region identified 13, 7, and 12 modules within the control, AsymAD and AD groups respectively, while the analysis in the EC identified 8, 8 and 11 modules within the control, AsymAD and AD groups respectively. GSEA analysis was performed for all fifty-nine modules to identify potential biological pathways the co-expressed genes may be involved with. A summary of the GSEA results on the co-expression module in the FC and EC is provided in Table 2.2 and Table 2.3 respectively.

2.3.3.1 Co-expression modules are weakly preserved in AsymAD and AD entorhinal cortex

Module preservation statistics were calculated for each brain region to identify co-expression networks that are weakly preserved through the course of the disease. Modules below a "preservation Zsummary" statistic of 10 and "preservation median rank" higher than the gold module (random 100 genes) are suggested to be weakly preserved. Module colours for the AsymAD and AD groups were mapped to the control module colours, allowing for changes and preservation in the co-expression

networks to be observed as the disease progresses. The module colours assigned in the EC brain region are independently assigned to modules colours assigned in the FC brain region and therefore, similar module colours across these two tissues bare no relation.

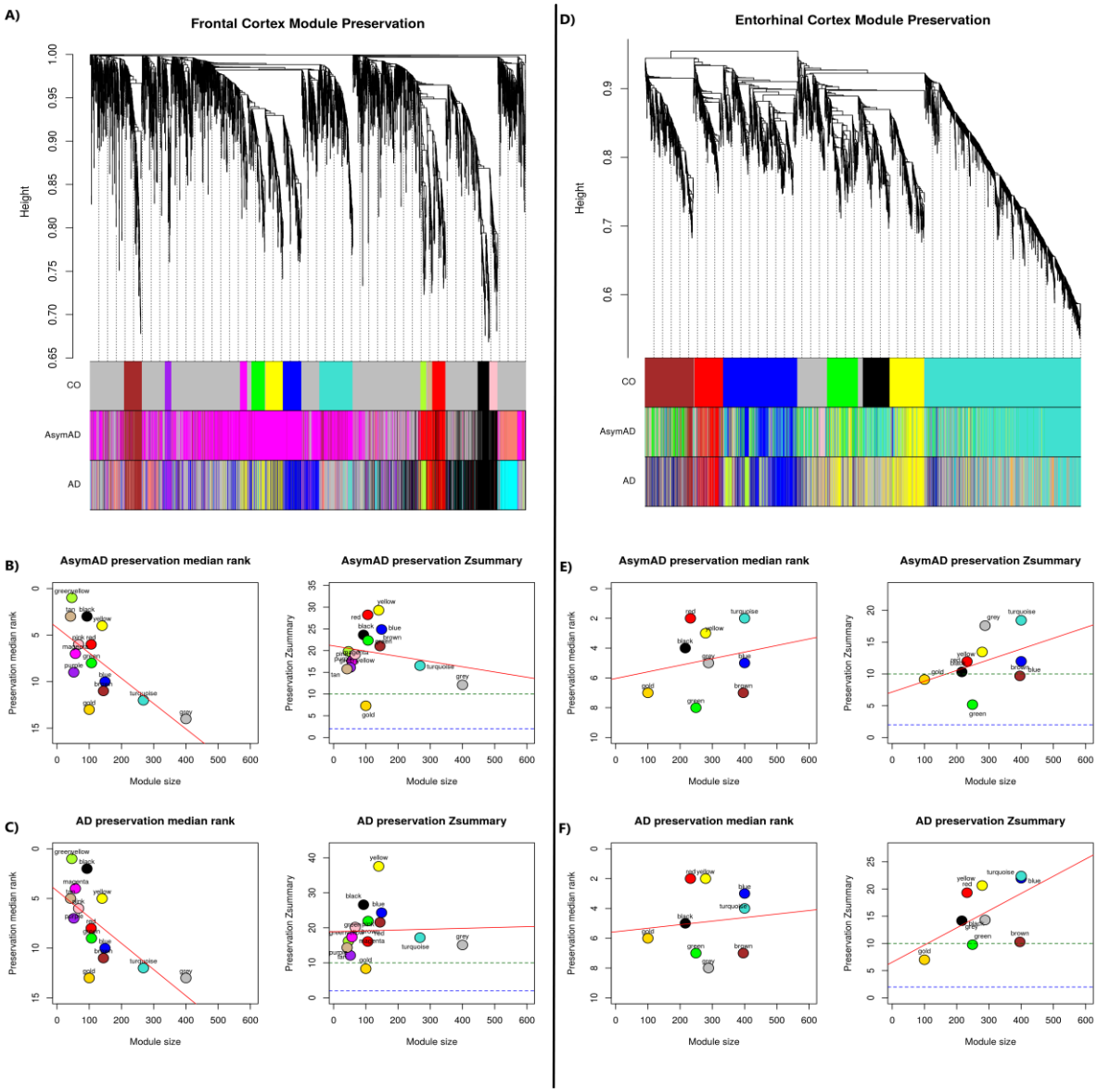


Figure 2.5: Hierarchical clustering of genes and module preservations statistics for the frontal cortex is illustrated in A-C) and entorhinal cortex in D-F). In brief, a co-expression network based on "signed" adjacency was independently created for all three phenotypes (control, AsymAD and AD group), topological overlap calculated, and hierarchical clustering used to group genes into modules. For the Hierarchical clustering plots, the y-axis represents the network distance with values closer to 0 indicating greater similarity of probe expression across the control group. The x-axis represents the modules in the control, AsymAD and AD group. The AsymAD and AD module colours are mapped to the control group, with the AsymAD and AD colour panel representing how well the control modules are preserved through the disease. The red line in the module preservation statistics (B, C, E, F) represents the correlation between module size and preservation statistics. The gold module represents 100 random genes, and the grey module represents uncharacterised genes. The FC preservation plots (B and C) suggest all modules in the control group are relatively preserved in the AsymAD and AD group. In contrast, the EC preservation plots (E and F) suggest the green module is not well preserved in the AsymAD and AD group and requires further investigation.

The FC "preservation Zsummary" statistics (Figure 2.5b and Figure 2.5c) suggests all modules from the control group are relatively well-preserved in the AsymAD and AD groups. In contrast, the EC "preservation median rank" statistics suggest the green control module is weakly preserved in AsymAD group (Figure 2.5e), and both the green and brown control modules are weakly preserved in the AD group (Figure 2.5f). In addition, the cross-tabulation statistics are also indicative of disruption to the EC green control module (Figure 2.6d). GSEA reveals the EC brown module in control, AsymAD and AD group is most significantly enriched for "selenocysteine synthesis" (control q-value=4.71e54, AsymAD q-value=5.89e-90, AD= 1.35E-96), suggesting this process is not significantly disrupted in AsymAD or AD subjects. In contrast, the EC control green module is significantly enriched (before multiple corrections) for **"neutrophil degranulation"** (p-value = 0.5e-4), **"TYROBP casual network"** (p-value = 2.5e-3) and the **"innate immune system"** (p-value = 2.7e-3), none of which are present in the green module of the AsymAD, suggesting these pathways may be disrupted in AsymAD subjects.

Clusters of co-expressed genes in both the FC and EC brain regions were enriched for specific cell types including neurons, astrocytes, oligodendrocytes and microglia (results not shown); however, a disturbance in any cell type in AsymAD subjects was not detected.

2.3.3.2 Frontal Cortex Co-expression network re-wired in AsymAD

Co-expression analysis identified 13 and 12 co-expressed modules in the control and AD subjects respectively. However, the AsymAD group exhibits 7 larger modules of highly co-expressed genes, suggesting the co-expression network is re-wired in the FC brain region in this intermediate stage of AD. The module preservation analysis suggested all modules within the control group are relatively preserved through the course of the disease, however, through cross-tabulation of the modules, subtle changes across individual modules can be seen to contribute to a much larger magenta module in the AsymAD group. The biological processes associated with the magenta module changes from being enriched for "glucose metabolism" (q-value 6.26e-02) in the control group to "oxidative phosphorylation" (q-value = 2.26e-11), Parkinson's disease (q-value = 5.12e-9), electron transport chain (q-value = 5.83e-9) and Alzheimer's disease (q-value = 8.21e-9) in the AsymAD group. Then the large magenta module in the AsymAD group, branches into four new AD modules (blue, turquoise, midnightblue, and yellow), which are most enriched for Parkinson's disease (q-value = 3.09e-4), neurotransmitter receptors and postsynaptic signal transmission (q-value = 0.01), the citric acid (TCA) cycle and respiratory electron transport (q-value = 0.01), and fas signalling (q-value=0.0030) respectively.

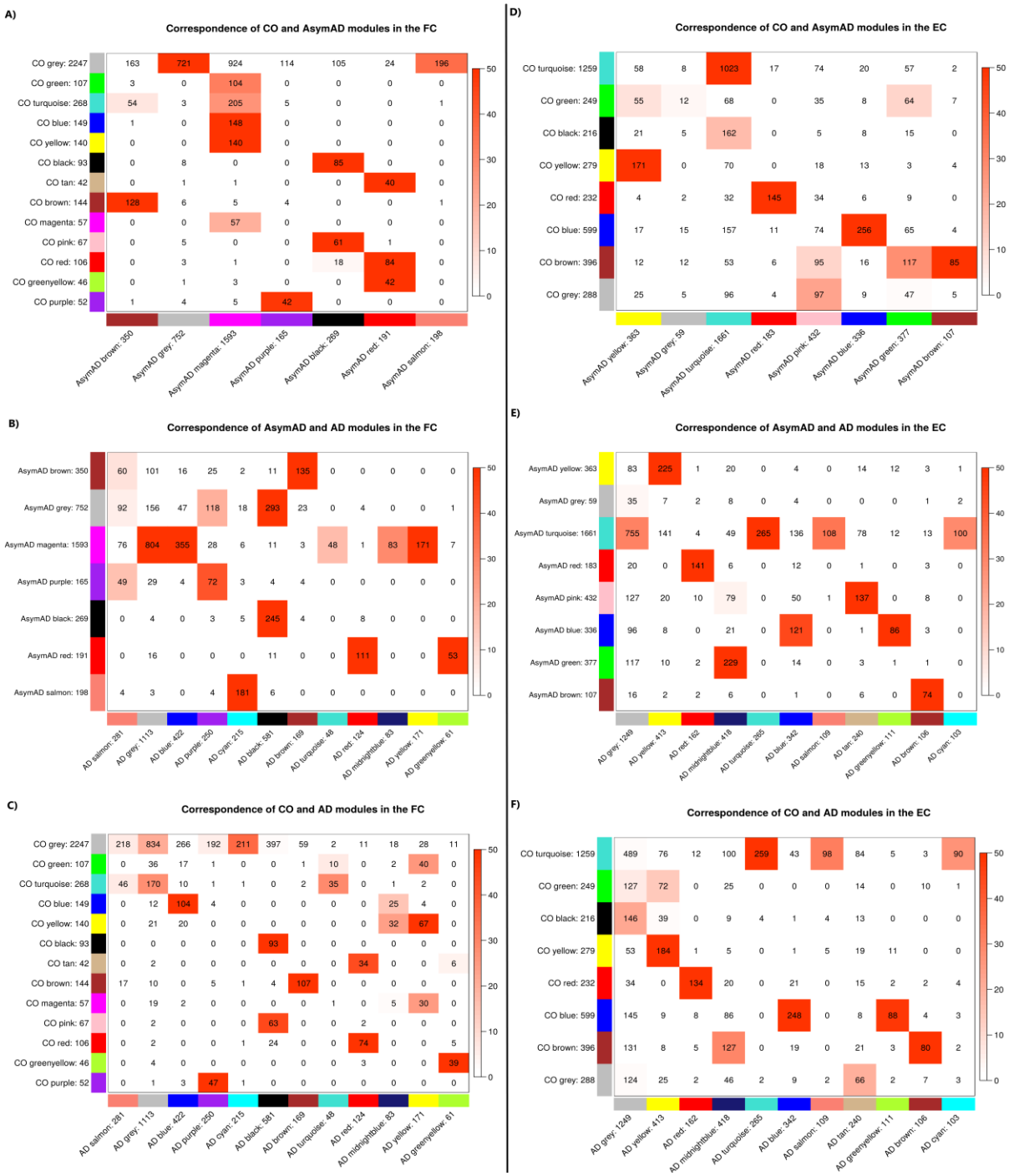


Figure 2.6: Illustrates the "module correspondence" between A) FC control and AsymAD group, B) FC AsymAD and AD group, C) FC control and AD group, D) EC control and AsymAD, E) EC AsymAD and AD group, and F) EC control and AD groups. The modules represent clusters of highly correlated genes which were calculated independently in each brain region and diagnosis group. The module colours in the AsymAD and AD group were assigned based on the gene overlap of the control module. The total number of genes within each module is indicated next to the module colour. The numbers in each cell represent the overlap of genes between modules, with increased red intensity cells indicating increased significant overlap based on Fisher's exact test. This "module correspondence" plot provides a visual overview of how modules of highly correlated genes are preserved or disrupted between, control, AsymAD and AD groups. Module preservation statistics suggested the green module in the EC control group is not well preserved in the AsymAD and AD groups, indicating possible disruption to the co-expression network in this module. This "module correspondence" plot identifies the disrupted genes in the control green module synchronises with the genes of the AsymAD yellow module, identifying the yellow module for further investigation.

2.3.3.3 Entorhinal Cortex yellow module enriched for all "Early AD" analysis DEG's

The yellow module contained all genes identified as significantly DE in the "Early AD" analysis (**ALDH2**, **FBLN1** and **METTL7A**) and contained a large number of genes disrupted from the green module, which was the least preserved module through disease progression. Overall, this made the yellow module a prime candidate for further investigation. Gene set enrichment analysis of the yellow module in the AsymAD group reveals enrichment in "**fatty acid degradation**" (q-value=0.03), "**glycerophospholipid metabolism**" (q-value=0.008), "**urea cycle and metabolism of arginine, proline, glutamate, aspartate and asparagine**" (q-value=0.05), "**astrocytic glutamate-glutamine uptake and metabolism**" (q-value=0.05) and "**neurotransmitter uptake and metabolism in glial cells**" (q-value=0.05), all of which were not previously enriched in the matched yellow module in the control group.

Protein-protein interaction analysis in the yellow control module generated six networks, with the largest containing 28 nodes and 30 edges, and identified **EGFR** gene as the only significant key hub gene (p-value=0.01). **APOE** was not a member of this network. Further PPI analysis in the AsymAD yellow module generated five networks, with the largest containing 71 nodes and 81 edges, and **EGRF** was still the key hub gene (p-value=0.007). In the equivalent AD yellow module, PPI analysis identified a single network generated with 284 nodes and 420 edges. This network contained far higher numbers of genes and now integrated the **APOE** gene as part of the network with **UBC** as the key hub (p-value=4.12e63). This suggests protein interactions in this yellow module increases gradually through the course of the disease, with up-regulated **EGRF** interacting with more genes in the AsymAD group when compared to controls, followed by significant changes occurring in the AD group where up-regulated **UBC** gene takes more of a central role.

Table 2.2: Summary of frontal cortex co-expression module GSEA results

Phenotype	Module	Module size	Most significant GSEA result	Pathway source	FDR adjusted q-value
CO	Black	93	Pentose phosphate pathway	HumanCyc	1.90E-02
	Blue	149	Parkinson's disease	KEGG	8.32E-10
	Brown	144	Differentiation Pathway	Wikipathways	2.26E-02
	Green	107	Oxidative phosphorylation	KEGG	1.43E-02
	GreenYellow	46	TNFs bind their physiological receptors	Reactome	1.60E-02
	Grey	2247	Processing of Capped Intron-Containing Pre-mRNA	Reactome	1.47E-02
	Magenta	57	Glucose metabolism	Reactome	6.26E-02
	Pink	67	Tight junction interactions	Reactome	1.49E-01
	Purple	52	Selenocysteine synthesis	Reactome	8.66E-53
	Red	106	TNF receptor superfamily (TNFSF) members mediating non-canonical NF-kB pathway	Reactome	3.88E-02
	Tan	42	Ovarian steroidogenesis	KEGG	4.43E-02
	Turquoise	268	Attenuation of gpcr signaling	BioCarta	6.13E-02
	Yellow	140	FCERI mediated MAPK activation	Reactome	5.62E-02
AsymAD	Black	269	RNA Polymerase II Transcription	Reactome	5.89E-04
	Brown	350	Differentiation Pathway	Wikipathways	3.17E-01
	Grey	752	mRNA Processing	Wikipathways	2.03E-02
	Magenta	1593	Oxidative phosphorylation - Homo sapiens (human)	KEGG	2.26E-11
	Purple	165	Peptide chain elongation	Reactome	5.03E-58
	Red	191	TNFR2 non-canonical NF-kB pathway	Reactome	2.71E-02
	Salmon	198	Hematopoietic cell lineage - Homo sapiens (human)	KEGG	2.82E-01
AD	Black	581	RNA Polymerase II Transcription	Reactome	6.61E-06
	Blue	422	Parkinson's disease	KEGG	3.09E-04
	Brown	169	Differentiation Pathway	Wikipathways	3.23E-02
	Cyan	215	Hematopoietic cell lineage - Homo sapiens (human)	KEGG	3.54E-01
	GreenYellow	61	Steroid hormone biosynthesis - Homo sapiens (human)	KEGG	2.82E-02
	Grey	1113	Neuronal System	Reactome	2.89E-01
	MidnightBlue	83	The citric acid (TCA) cycle and respiratory electron transport	Reactome	1.31E-02
	Purple	250	Eukaryotic Translation Elongation	Reactome	9.36E-66
	Red	124	TNF receptor superfamily (TNFSF) members mediating non-canonical NF-kB pathway	Reactome	9.31E-02
	Salmon	281	Histidine metabolism	EHMN	4.87E-04
	Turquoise	215	Neurotransmitter receptors and postsynaptic signal transmission	Reactome	1.15E-02
	Yellow	171	Fas	INOH	2.94E-03

Co-expression analysis in the frontal cortex brain region identified 13, 7, and 12 modules within the control, AsymAD and AD groups respectively. Gene set enrichment analysis was performed on each module, and the most significant result from each module is provided above.

Table 2.3: Summary of entorhinal cortex module co-expression results

Phenotype	Module	Module size	Most significant GSEA result	Pathway source	FDR adjusted q-value
CO	Black	216	Hedgehog	INOH	7.69E-03
	Blue	599	Generic Transcription Pathway	Reactome	3.32E-05
	Brown	396	Ribosome	KEGG	4.71E-54
	Green	249	Neutrophil degranulation	Reactome	1.35E-01
	Grey	288	Pink/Parkin Mediated Mitophagy	Reactome	2.97E-01
	Red	232	Leptin Insulin Overlap	Wikipathways	6.01E-02
	Turquoise	1259	Neuronal System	Reactome	1.68E-08
	Yellow	279	Histidine metabolism	EHMN	5.24E-03
AsymAD	Blue	336	Generic Transcription Pathway	Reactome	2.04E-04
	Brown	107	Eukaryotic Translation Elongation	Reactome	1.13E-93
	Green	337	Processing of Capped Intron-Containing Pre-mRNA	Reactome	2.08E-06
	Grey	1661	Antigen processing and presentation	KEGG	2.10E-04
	Pink	432	Neural Crest Differentiation	Wikipathways	1.44E-01
	Red	183	Hematopoietic cell lineage	KEGG	2.21E-01
	Turquoise	1661	Parkinson's disease	KEGG	2.82E-10
	Yellow	363	Metallothioneins bind metals	Reactome	8.08E-03
AD	Blue	342	Generic Transcription Pathway	Reactome	6.31E-04
	Brown	106	Eukaryotic Translation Elongation	Reactome	4.46E-98
	Cyan	103	Cardiac conduction	Reactome	4.69E-04
	GreenYellow	111	Steroid hormone biosynthesis - Homo sapiens (human)	KEGG	2.04E-01
	Grey	1249	Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.	Reactome	3.55E-06
	MidnightBlue	418	Processing of Capped Intron-Containing Pre-mRNA	Reactome	7.92E-07
	Red	162	Fat digestion and absorption	KEGG	2.80E-01
	Salmon	109	How progesterone initiates the oocyte maturation	BioCarta	2.08E-02
	Tan	240	Differentiation Pathway	Wikipathways	1.95E-01
	Turquoise	265	Neuronal System	Reactome	5.10E-11
	Yellow	413	Propanoate metabolism	EHMN	3.67E-03

Co-expression analysis in the entorhinal cortex brain region identified 8, 8, and 11 modules within the control, AsymAD and AD groups respectively. Gee set enrichment analysis was performed on each module, and the most significant result from each module is provided above.

2.4 Discussion

2.4.1 Transcriptomic perturbations suggest AsymAD subjects could be an intermediate stage between control and MCI/AD

This study hypothesises the samples labelled as "AsymAD" subjects are an intermediate state between healthy ageing and MCI/AD. The assignment of these samples to the AsymAD group was based on the fact that these individuals had no reported clinical record of dementia prior to death as indicated in the MRC-LBB database; however, upon autopsy, these samples were found to have low levels of hallmark AD pathology, i.e. BRAAK Staging >= 2. Furthermore, an independent expression study identified the TRIL gene as being highly correlated with AD neuropathology, specifically tau pathology (Miyashita *et al.*, 2014). This study shows that the TRIL gene expression gradually increases from the Control to AsymAD, and then further increases in AD subjects, and this expression pattern is only observed in brain regions known to be affected by hallmark AD pathology (amyloid and NFT's), i.e. the EC, TC and FC, and not in the CB brain region. This observation suggests the phenotype assignments (controls, AsymAD, AD) are a suitable representation of three points in AD progression (assuming the AsymAD subjects are all prodromal AD), and as suggested by the TRIL gene expression pattern across brain regions and the fact the CB has been consistently reported to be partially spared from hallmark AD pathology (amyloid and NFT's), even those with severe AD pathology (Convit *et al.*, 2000), genes whose expression pattern differs significantly in the CB from that consistently seen in the EC, TC and FC tissues may be associated with hallmark AD pathology.

2.4.2 MOSPD3 gene is perturbed in the brains of AsymAD and blood of AD subjects.

The MOSPD3 gene was the only significant DE gene which is consistently down-regulated across all four brain regions in the AsymAD subjects suggesting this may be an early marker of cell dysfunction in AD. The MOSPD3 gene encodes for a Motile Sperm Domain Containing 3 protein [provided by RefSeq, Jul 2008] and has been reported to be significantly down-regulated (p-value = 6.47E-05) in the blood of AD subjects when compared to MCI subjects (Lunnon *et al.*, 2012). This suggests MOSPD3 gene expression is significantly decreased in the brain before clinical signs of AD are apparent; however, blood gene expression levels are only significantly decreased after clinical signs of AD are apparent. It is difficult to interpret the biological relevance of this gene in AD, and further investigation is required.

2.4.3 Genes perturbed in brain regions affected explicitly by hallmark AD pathology may be associated with plaques and tangles, providing new therapeutic targets.

Many molecular and cellular changes occur in AD brains including nerve cell death, atrophy, loss of neurons and accumulation hallmark AD pathology, specifically plaques and tangles. However, not all brain regions are affected to the same degree. The CB, which only accounts for 10% of the brain but

contains over 50% of the brains total neurons, is often regarded as being partially spared from AD as plaques and tangles are generally not reported (Convit *et al.*, 2000; Jacobs *et al.*, 2017), and in this study are free from hallmark AD pathology in both AsymAD and AD subjects. For subjects with hallmark AD pathology (BRAAK >=2, AsymAD and AD), genes significantly and consistently perturbed across the EC, TC and FC tissues that are not or are significantly reversed in the CB, may be associated with hallmark AD pathology, although, it remains unclear if these genes are causative or a response to the pathology itself.

This study identified a total of 15 genes (ALDH2, FBLN2, METTL7A, FLCN, ASPHD1, ARL5A, GPR162, HBA2, PCID2, NDRG2, BEND3, RAP1Gap, GPM6B, ANKEF1 and NPC2) with expression patterns suggestive of an association with hallmark AD pathology. Previous studies have already demonstrated an increased expression of ALDH2 accelerated neurodegeneration and increased the accumulation of hyperphosphorylated tau protein (Ohsawa *et al.*, 2008) in mice, while another demonstrated NDRG2 might play a role in generating Aβ (Rong *et al.*, 2017). Collectively, the 15 genes are not significantly enriched to be involved with any biological pathway; however, individually, these genes may play an essential role in the pathological aspect of AD and may provide new therapeutic targets for disease intervention.

2.4.4 Individuals with milder disease (early BRAAK pathology) show increased changes in the frontal cortex compared to the entorhinal cortex.

The molecular changes in AD may initially begin in the FC, a region involved in working memory, as there were relatively more changes in the FC of mild pathology AD cases (AsymAD) than the EC region. This mirrors changes described in a longitudinal study involving ageing controls, where positron emission tomography (PET) scans were used to detect increased activity in the medial frontal cortex and decrease activity in the temporal lobe brain region in subjects who subsequently acquired cognitive impairment (Beason-Held *et al.*, 2013). In addition, a higher degree of atrophy has also been detected in the FC than the temporal lobe brain region in MCI when compared to AD (Tabatabaei-Jafari, Shaw and Cherbuin, 2015). This study provides further evidence to suggest that brain perturbations at the molecular/transcriptomic level may initially occur in the FC before the presentation of more severe clinical symptoms consistent with a diagnosis of probable AD.

At the later point of the disease when clinical signs of AD are present, most substantial number of transcriptomic changes is detected in the EC, followed by the TC, FC and only minor changes in the CB. This observation matches the common route AD neuropathology is seen to spread through the brain. Furthermore, more DEG in the “Late AD” analysis is detected when compared to “Early AD” analysis,

signifying more genes are disrupted in the later stage of the disease when the clinical symptoms of cognitive impairment are apparent.

2.4.5 Neutrophil, TYROBP network and the innate immune system disrupted in Asymptomatic AD

Co-expression analysis of the EC brain region identified a green module of highly co-expressed genes, which is disrupted in the AsymAD and AD subjects according to both module preservation statistics and cross-tabulation analysis. This green module is significantly enriched for "**neutrophil degranulation**", "**TYROBP casual network**" and the "**innate immune system**" processes in the control subjects, but not in the AsymAD or AD subjects, suggesting these pathways are most likely disrupted during the disease. Disturbance in TYROBP and Immune system pathways have been widely accepted in AD (Lambert *et al.*, 2010; Ma *et al.*, 2015), and a previous mouse study demonstrated disruptions in neutrophil levels impact memory loss and neurological features of AD (Pietronigro *et al.*, 2017). This study now suggests these pathways are specifically perturbed in the EC brain region early in the disease when hallmark AD pathology exists but clinical symptoms of AD are absent.

2.4.6 Disruption in brain energy pathways is detectable early in the disease

Co-expression analysis of the FC identifies disruptions in the "**glucose metabolism**", "**glucogenesis**" and"**oxidative phosphorylation**" processes in the AsymAD group, while DE analysis identified disruption in the "**gluconeogenesis and glycolysis**" pathway in the AD subjects. The brain critically relies on a constant supply of energy which is known to be generated by glycolysis followed by oxidative phosphorylation. Changes in the brain energy pathways have been widely accepted in AD (Shoffner, 1997; Cunnane *et al.*, 2011), with a general decrease in glycolysis suggested to be a result of decreased brain functionality. This study demonstrates disruptions in the energy pathway are detectable early in the disease, in subjects with low levels of AD pathology.

2.4.7 The Glutamate-Glutamine Cycle is disturbed in AsymAD and AD subjects

Gene set enrichment analysis on DEGs identified the "**glutamate-glutamine cycle**" as the only biological pathway significantly perturbed across all brain regions in the AsymAD subjects. Furthermore, co-expression analysis of the EC brain regions was indicative of disruptions to the "**urea cycle and metabolism of arginine, proline, glutamate, aspartate and asparagine**" and "**astrocytic glutamate-glutamine uptake and metabolism**" in AsymAD and AD subjects, further confirming a possible disruption in glutamate-related activities in the brain.

Astrocytes are the most common form of neuroglial cells in the brain, and its primary function is to protect neurons against excitotoxicity by converting excess ammonia and glutamate to glutamine through the glutamate-glutamine cycle. Glutamate is the principal excitatory neurotransmitter in the

brain and plays a vital role in linking carbohydrate and amino acid metabolism via the tricarboxylic acid (TCA) cycle. Glutamate is also a precursor of γ-aminobutyric acid (GABA) which binds and inhibits neuron activity; hence, an accumulation of glutamate can cause failures in synaptic connectivity, leading to deficient cognition and memory (Schousboe *et al.*, 2014). A disruption in the glutamate-glutamine cycle would have a severe knock-on effect on many other biological pathways, including a disruption in amino acid metabolism which could explain the enrichment of "urea cycle and metabolism of arginine, proline, glutamate, aspartate and asparagine" in the results as well. In addition, glutamate stimulates astrocytes to derive energy from oxidative and glycolytic pathways, both of which have been identified as disrupted in AsymAD subjects.

The genes enriched in this pathway were all significantly up-regulated, indicating an overactive cycle. This could be part of the brain defence mechanism in preventing the accumulation of brain glutamate levels or a broken cycle which is consistently being overactive, leading to decreased levels of brain glutamate, a phenomenon observed in AD subjects (Fayed *et al.*, 2011). Targetting this pathway for AD treatment is extraordinarily complex and challenging as over inhibition or excitation may lead to increased levels of glutamate and glutamine respectively, both of which can be neurotoxic at high levels. Therapeutic compounds affecting the "glutamate-glutamine cycle" have already been identified, such as memantine, which is already a clinically established therapeutic drug used to for the symptomatic treatment of AD, which blocks N-methyl-d-aspartate (NMDA) receptors (Johnson and Kotermanski, 2006), essentially preventing excitotoxicity caused by neurotransmitters such as glutamate and ultimately increasing cognition temporarily.

The glutamate-glutamine cycle has been previously suggested to be disrupted in AD (Walton and Dodd, 2007), along with many other central nervous system disorders including Huntington's disease and Amyotrophic Lateral Sclerosis (ALS) (Matthews, Henter and Zarate, 2012). This study demonstrates this is one of the earliest biological pathways perturbed across all brain regions in AD, before clinical symptoms of AD are apparent, which can have a knock-on effect on other biological pathways also observed to be disrupted in the disease. Clinically established drugs to relieve AD symptoms already interact with this pathway and could also be effective in the asymptomatic period to prolong cognitive impairment, although clinical identification and measuring effectiveness in AsymAD subjects would be a challenge in itself.

2.4.8 Co-expression network changes indicate a shift from "cell proliferation" in AsymAD subjects to "removal of amyloidogenic proteins" in AD subjects.

Protein-protein interactions identified EGFR as a key hub gene in both the control and AsymAD groups; however, it achieves more connections with neighbouring proteins in the AsymAD group, suggesting a

possible increase in the EGFR activity. The EGFR gene is up-regulated in the AsymAD group and encodes for a transmembrane glycoprotein that binds to epidermal growth factor, leading to cell proliferation. In contrast, EGFR is replaced by UBC as the key hub gene in AD subjects, indicating it may play a more central role in the disease once the accumulation of hallmark AD pathology is at a level where clinical symptoms are apparent. The UBC gene is significantly up-regulated in the EC of AD subjects and is considered a stress gene which encodes for polyubiquitin precursor protein, a member of the ubiquitin-proteasome system (UPS) which removes toxic proteins and impacts on the amyloidogenic pathway of amyloid precursor protein (APP) processing that generates Abeta (Hong, Huang and Jiang, 2014). A previous AD study had also observed UBC as a novel key hub gene and demonstrated UBC knockout models in C. elegans accelerated age-related AB toxicity (Mukherjee *et al.*, 2017). Effectively, a portion of the co-expression network may have a central role involved in cell proliferation in control subjects, with increased activity in AsymAD subjects, followed by a shift towards the removal of toxic proteins such as amyloid-beta in AD subjects.

2.4.9 Limitations

Several individual processes are involved during sample preparation and during microarray gene expression, including, transcriptional steps, labelling, hybridisation and intensity measurement. Each of these steps can introduce technical variations (Richard *et al.*, 2014). Gene expression microarrays are also affected by non-biological variations, such as reagent lot numbers, different technicians and even atmospheric ozone levels (Chen *et al.*, 2011). In addition, the quantification of gene expression in post-mortem brains has been shown to be complicated by confounding factors such as gender, age at death and brain pH (Preece and Cairns, 2003). It is impossible to measure and account for all variables that may be influencing how genes are expressed, however, a study has demonstrated SVA can predict and account for these latent effects (Leek and Storey, 2007). Although SVA was incorporated to account for hidden batch effects, it does not guarantee that all technical variation is completely removed.

This study cannot exclude the fact AsymAD group may represent a heterogeneous group consisting of cognitively normal, MCI, mixed dementia and AD subjects. It remains unclear these AsymAD subjects would remain free from clinical symptoms of dementia with longer survival and can be argued to be a possible extension to general ageing. However, the extent of BRAAK staging in AsymAD subjects was at a level consistently found with early cognitive impairment, and therefore, this study makes the strong assumption that these subjects are more likely to be prodromal AD rather than an extension of natural ageing. As AsymAD subjects are extremely rare, hence the low sample numbers in this study, larger AsymAD cohorts are required for better discovery and to validate the current findings.

2.5 Conclusion

This is the first study to explore the emergence of transcriptomic changes in the human brain from normal ageing through to mild AD pathology and diagnosis of AD. Using DE analysis, coupled with a "systems biology" approach, this study was able to detect disturbances in the energy pathways and the "glutamate-glutamine cycle" in the brains of subjects with mild and severe AD pathology. This study found that changes in the FC brain region dominate in mild pathology, but are greater in the EC in subjects with more severe pathology, thus mirroring the changes in the aggregate spread in AD. This study provides new insight into the earliest biological changes occurring in the brain prior to AD diagnosis while providing new potential therapeutic targets.

Chapter 3: Analysis of Alzheimer's Disease brain transcriptomic data (Article)

Journal of Alzheimer's Disease 68 (2019) IN PRESS
DOI 10.3233/JAD-181085
IOS Press

A Meta-Analysis of Alzheimer's Disease Brain Transcriptomic Data

Hamel Patel [a, b], Richard J.B Dobson [a, b, c, d, e, *] Stephen J Newhouse [a, b, c, d, e, *]

[a] *Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK*

[b] *NIHR BioResource Centre Maudsley, NIHR Maudsley Biomedical Research Centre (BRC) at South London and Maudsley NHS Foundation Trust (SLaM) & Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London.*

[c] *Health Data Research UK London, University College London, 222 Euston Road, London, UK*

[d] *Institute of Health Informatics, University College London, 222 Euston Road, London UK*

[e] *The National Institute for Health Research University College London Hospitals Biomedical Research Centre, University College London, 222 Euston Road, London, UK*

Accepted 12th February 2019

Abstract.

Background: Microarray technologies have identified imbalances in the expression of specific genes and biological pathways in Alzheimer's disease (AD) brains. However, there is a lack of reproducibility across individual AD studies, and many related neurodegenerative and mental health disorders exhibit similar perturbations.

Objective: Meta-analyse publicly available transcriptomic data from multiple brain-related disorders to Identify robust transcriptomic changes specific to AD brains.

Methods: Twenty-two AD, eight Schizophrenia, five Bipolar Disorder, four Huntington's disease, two Major Depressive Disorder and one Parkinson's disease dataset totalling 2667 samples and mapping to four different brain regions (temporal lobe, frontal lobe, parietal lobe and cerebellum) were analysed. Differential expression analysis was performed independently in each dataset, followed by meta-analysis using a combining p-value method known as Adaptively Weighted with One-sided Correction.

Results: Meta-analysis identified 323, 435, 1023 and 828 differentially expressed genes specific to the AD temporal lobe, frontal lobe, parietal lobe and cerebellum brain regions respectively. Seven of these genes were consistently perturbed across all AD brain regions with SPCS1 gene expression pattern replicating in RNA-Seq data. A further nineteen genes were perturbed specifically in AD brain regions affected by both plaques and tangles, suggesting possible involvement in AD neuropathology. In addition, biological pathways involved in the "metabolism of proteins" and viral components were significantly enriched across AD brains.

Conclusion: This study identified transcriptomic changes specific to AD brains, which could make a significant contribution towards the understanding of AD disease mechanisms and may also provide new therapeutic targets.

KEYWORDS: Alzheimer's disease, Mental disorders, Neurodegenerative disorders, Microarray analysis, Gene expression, Human, Neuropathology, Meta-analysis

*These authors are joint last authors.
Correspondence to: Hamel Patel, Department of Biostatistics & Health Informatics, SGDP Centre, IoPPN, De Crespigny Park, Denmark Hill London, SE5 8AF, UK. Tel.: +44 207 848 0969; E-mail: hamel.patel@kcl.ac.uk.

INTRODUCTION

Alzheimer's disease (AD) is the most common form of dementia affecting over 44 million individuals worldwide, and numbers are

expected to triple by 2050 [1]. The hallmark of the disease is characterised by the abnormal brain accumulation of amyloid-β (Aβ) protein and hyperphosphorylated tau filaments, which forms structures known as plaques and tangles respectively. The accumulation of these proteins contributes to the loss of connections between neurone synapses, leading to the loss of brain tissue and the disruption of normal cognitive functions.

As AD progresses, the spread of plaques and tangles in the brain usually occurs in a predictable pattern and can begin up to 18 years prior to the onset of clinical symptoms [2]. In the earliest stages of the disease, plaques and tangles form in areas of the brain primarily involved in learning and memory, specifically the hippocampus and entorhinal cortex, both situated in the temporal lobe (TL) region [3]. Next, the frontal lobe (FL), a region involved in voluntary movement, is affected, followed by the parietal lobe (PL), a region involved in processing reading and writing. In the later stage of the disease, the occipital lobe, a region involved in processing information from the eyes, can become affected, followed by the cerebellum (CB), a region which receives information from the sensory systems and the spinal cord to regulates motor movement. Nerve cell death, tissue loss and atrophy occur throughout the brain as AD progresses, leading to the manifestation of clinical symptoms associated with loss of normal brain function. However, not all brain regions are neuropathologically affected in the same manner. The CB, which only accounts for 10% of the brain but contains over 50% of the brains total neurones, is often neglected in AD research because it is generally considered to be partially spared from the disease as plaques are only occasionally seen but tangles are generally not reported [4,5].

The histopathological spread of the disease is well documented, and with the advent of high throughput genomics approaches, we are now able to study the transcriptomic and biological pathways disrupted in AD brains. Microarrays can simultaneously examine thousands of genes, providing an opportunity to identify imbalances in the expression of specific genes and biological pathways. However, microarray reproducibility has always been questionable, with replication of differentially expressed genes (DEG's) very poor [6]. For example, two independent microarray transcriptomic studies performed differential expression analysis in the hippocampus of AD brains. The first study by Miller et al. identified 600 DEG's [7], and a similar study by Hokama et al. identified 1071 DEG's [8]. An overlap of 105 DEG's exist between the two studies; however, after accounting for multiple testing, no gene was replicated between the two studies. The Miller study consisted of 7 AD and 10 control subject's expression profiled on the Affymetrix platform while the Hakoma study consisted of 31 AD and 32 control subjects expression profiled on the Illumina platform. Replication between the Illumina and Affymetrix platform has been shown to be generally very high [9]; therefore, the lack of replication between the two studies is probably down to a range of other factors including low statistical power, sampling bias and disease heterogeneity.

Unlike DEG's, replication of the molecular changes at a pathway level are more consistent and have provided insights into the biological processes disturbed in AD. Numerous studies have consistently highlighted disruptions in immune response [10–13] protein transcription/translation [10,11,14–17], calcium signalling [10,18,19], MAPK signalling [7,16] various metabolism pathways such as carbohydrates [16], lipids [16,20], glucose [17,21,22], and [11,23], chemical synapse [7,18,19] and neurotransmitter [11,18,19] However, many of these pathways have also been suggested to be disrupted in other brain-related disorders. For example, disruptions in calcium signalling, MAPK, chemical synapse and various neurotransmitter pathways have also been implicated in Parkinson's Disease (PD) [24,25] In addition, glucose metabolism, protein translation, and various neurotransmission pathways have also been suggested to be disrupted in Bipolar Disorder (BD[26–29]. Although the biological disruptions involved in AD are steadily being identified, many other neurodegenerative and mental disorders are showing similar perturbations.

We are yet to identify robust transcriptomic changes specific to AD brains.

In this study, we combined publicly available microarray gene expression data generated from AD human brain tissue and matched cognitively healthy controls to conduct the most extensive AD transcriptomic microarray meta-analyses known to date. We generate AD expression profiles across the temporal lobe, frontal lobe, parietal lobe and cerebellum brain regions. We further refine each expression profile by removing perturbations seen in other neurodegenerative and mental disorders (PD, BD, Schizophrenia [SCZ], Major Depressive Disorder [MDD] and Huntington's Disease [HD]) to decipher specific transcriptomic changes occurring in human AD brains. These AD-specific brain changes may provide new insight and a better understanding of the disease mechanism, which in turn could provide new therapeutic targets for preventing and curing AD.

MATERIALS AND METHODS

Selection of publicly available microarray studies

Publicly available microarray gene expression data was sourced from the Accelerating Medicines Partnership-Alzheimer's Disease AMP-AD (doi:10.7303/syn2580853, doi:10.1038/ng.305, doi:10.1371/journal.pgen.1002707, doi:10.1038/ng.305, doi:10.1038/sdata.2016.89, doi:10.1038/sdata.2018.185) and ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) in June 2016. For a study to be selected for inclusion, the data had to (1) be generated from a neurodegenerative or mental health disorder, (2) be sampled from human brain tissue, (3) have gene expression measured on either the Affymetrix or Illumina microarray platform, (4) contain both diseased and suitably matched healthy controls in the same experimental batch and (5) contain at least 10 samples from both the diseased and control group.

Microarray gene expression data pre-processing

Data analysis was performed in RStudio (version 0.99.467) using R (version 3.2.2). All data analysis scripts used in this study are available at https://doi.org/10.5281/zenodo.823256. In brief, raw Affymetrix microarray gene expression data was "mas5" background corrected using R package "affy" (version 1.42.3) and raw Illumina microarray gene expression data Maximum Likelihood Estimation (MLE) background corrected using R package "MBCB" (version 1.18.0). Studies with samples extracted from multiple tissues were separated into tissue-specific matrices, log2 transformed and then Robust Spline Normalised (RSN) using R package "lumi" (version 2.16.0).

BRAAK staging is a measure of AD pathology and ranges from I-VI. In general, stages I-II, III-IV and V-VI represent the "low likelihood of AD", "probable AD" and "definite AD" respectively [30]. To maintain homogeneity within the sample groups and to be able to infer pathological related genetic changes, if BRAAK staging was available, clinical AD samples with BRAAK scores ≤ 3 or clinical control samples with BRAAK scores ≥ 3 were removed from further analysis.

Gender was predicted using the R package "massiR" (version 1.0.1) and used to subset the data into four groups based on diagnosis (case/control) and gender (male/female). Next, probes below the 90th percentile of the log2 expression scale in over 80% of samples were deemed "not reliably detected" and were excluded from further analysis to eliminate noise [31] and increase power [32].

Publicly available data is often accompanied by a lack of sample processing information, making it impossible to adjust for known systematic errors introduced when samples are processed in multiple batches, a term often known as "batch effects". To account for both known and latent variation, batch effects were

estimated and removed using the Principal Component Analysis (PCA) and Surrogate Variable Analysis (SVA) using the R package "sva" (version 3.10.0). Gender and diagnosis information were used as covariates in sva when correcting for batch effects. Outlying samples were iteratively identified and removed from each gender and diagnosis group using fundamental network concepts described in [33]. Platform-specific probe ID's were converted to Entrez Gene ID's using the BeadArray corresponding R annotation files ("hgu133plus2.db", "hgu133a.db", "hgu133b.db", "hugene10sttranscriptcluster.db", "illuminaHumanv4.db", "illuminaHumanv3.db") and differential expression analysis was performed within each dataset using the R package "limma" (version 3.20.9).

Finally, study compatibility analysis was investigated through the R package "MetaOmics" (version 0.1.13). This package uses differentially expressed genes (DEGs), co-expression and enriched biological pathways analysis to generate six quantified measures that are used to generate a PCA plot. The direction of each Quality Control (QC) measure is juxtaposed on top of the two-dimensional PC subspace using arrows. Datasets in the negative region of the arrows were classed as outliers [34] and were removed from further analysis.

Meta-analysis

Datasets were grouped by the primary cerebral cortex lobes (TL, FL, PL) and the CB. Meta-analysis was performed using a "combining p-values" method known as "Adaptively Weighted with One-sided Correction" (AW.OC), implemented through the R package "MetaDE" (version 1.0.5)[34]. A combining p-value method was chosen to address the biases introduced from different platforms. AW.OC was chosen as it permits missing information across datasets which are introduced by combining data generated from different microarray platforms and expression chips. This avoids the need to subset individual datasets to common probes, which essentially allows for the maximum number of genes to be analysed. Furthermore, the method provides additional information on which dataset is contributing towards the meta-analysis p-value, and has been shown to be amongst the best performing meta-analysis methods for combining p-values for biological associations [35]. The meta-analysis method does not provide an overall directional change for each gene; therefore, the standard error (SE) was calculated from the DE logFC values of each gene across the AW assigned significant datasets and used for standard meta-summary estimate analysis using the R package "rmeta" (version 2.16). This served as the "meta expression" change in downstream analysis where positive values represent a gene being up-regulated in AD and negative values as being down-regulated in AD. Selecting DE genes based on an arbitrary expression change significantly influences the interpretation of DE results [36]. At least half of differential expression based studies incorporate a fold change cut-off typically between 2-3, however, informative RNAs and expressed transcripts have been shown to have a fold change less than 2 [37], and genes with low fold change have been demonstrated to influence biological effects in signalling cascades and pathways [36]. In addition, gene expression is heavily influenced by tissue, and as this study performs meta-analysis across multiple inter-related tissues within larger brain compartments, we do not employ an arbitrary fold change cut-off to determine if a gene is differentially expressed, however, we do require the gene to be consistently expressed across these tissues. if a gene was significantly DE according to the meta-analysis (FDR adjusted meta p-value ≤ 0.05), but at least one contributing dataset (according to AW.OC weights) had directional logFC discrepancy (i.e. up-regulated in one dataset and down-regulated in another dataset), the gene was deemed to be discordant and was excluded from further analysis. This ensured we only captured robust, and consistently reproducible expression signatures.

Generation of disease-specific meta-analysis expression profiles

Meta-analysis was performed across all AD datasets, followed by a separate meta-analysis across the non-AD disorder datasets. Using these meta-analysis results we generated three expression profiles; (1) "AD expression profile", (2) "AD-specific expression profile" and (3) "common neurological disorder expression profile".

The first expression profile, "AD expression profile", is a direct result of the meta-analysis performed on AD studies, which represents the changes typically observed from an AD and cognitively healthy control study design. The second expression profile, deemed as the "AD-specific expression profile", is produced by subtracting significantly DEG's found in the non-AD meta-analysis results from the "AD expression profile". This profile represents transcriptomic changes specifically observed in AD and not in any other neurodegenerative or mental health disorder used in this study. The third expression profile, deemed as the "common neurological disorder expression profile", represents genes which are significantly DE in all disorders used in this study, including AD.

Replication of significant microarray genes in RNA-Seq data

The genes significantly DE and deemed to be of biological significance in this study were queried in the curated web-based database Agora (data version 9, accessible at https://agora.ampadportal.org), which provides expression change of genes in AD based on RNA-Seq of 2100 human brain samples.

Functional and gene set enrichment analysis

Gene set enrichment analysis (GSEA), and Gene Ontology (GO) analysis was conducted using an Over-Representation Analysis (ORA) implemented through the ConsensusPathDB web platform (version 32) [38] in May 2017. ConsensusPathDB incorporates numerous well-known biological pathway databases including BioCarta, KEGG, Reactome and Wikipathways. The platform performs a hypergeometric test while integrating a background gene list, which in this case is a list of all the genes that pass quality control in this study, compiles results from each database and corrects for multiple testing using the false discovery rate (FDR) [38]. A minimum overlap of the query signature and database was set to 2, and a result was deemed significant if the q-value was ≤ 0.05.

Network analysis

Protein-protein interaction (PPI) networks were created by uploading the meta-analysis DEG lists (referred to as seeds in network analysis) along with their meta logFC expression values to NetworkAnalyst's web-based platform http://www.networkanalyst.ca/faces/home.xhtml in June 2017. The "Zero-order Network" option was incorporated to allow only seed proteins directly interacting with each other, preventing the well-known "Hairball effect" and allowing for better visualisation and interpretation [39]. Sub-modules with a p-value ≤ 0.05 (based on the "InfoMap" algorithm [40]) were considered significant key hubs, and the gene with the most connections within this hub was regarded as the key hub gene.

RESULTS

The AD microarray datasets

We Identified and acquired nine publicly available AD studies from ArrayExpress and AMP-AD, of which seven studies contained samples extracted from differing regions of the brain. The basic characteristics of each study and dataset are provided in *Table 1.* Separating the nine studies by brain regions resulted in 46 datasets. Here a "dataset" is defined by brain region and study origin. For example, ArrayExpress study E-GEOD-36980 consists of diseased and healthy samples extracted from three different tissues (temporal cortex, hippocampus and frontal cortex). All samples originating from the same tissue were classified as one dataset; therefore, study E-GEOD-36980 generated three datasets, representing the three different tissues.

The 46 AD datasets contained both AD samples and healthy controls, were assayed using seven different expression chips over two different microarray platforms (Affymetrix and Illumina) and consisted of a total 2718 samples before QC. Briefly, the MetaOmics analysis identified study syn4552659 as an outlier and was therefore removed from further analysis (see supplementary text 1), resulting in 1501 samples (746 AD, 755 controls) in the remaining 22 datasets after QC

Summary of the AD meta-analysis DEG counts

The AD meta-analysis was performed on the 22 AD datasets and independently identified differentially expressed genes within the TL, FL, PL and CB brain regions. A summary of the number of datasets in each brain region and the number of significant DEG's identified is provided in *Table 2. The* complete DE results are provided in *Supplementary Table 1*. As mentioned in the methods, due to gene expression being influenced by tissue source and as this study incorporates different brain regions, we do not employ an arbitrary cut-off value to determine genes that are highly or lowly expressed, but primarily focus on genes consistently perturbed. However, we provide the meta expression values in supplementary tables and advise readers to consider the expression distribution of all DEG's within each brain region independently, if determining whether a gene is highly/lowly expressed. For instance, the cerebellum meta expression ranges from -0.53 to 0.54 with an interquartile range(Q1-Q3) of -0.1 to 0.1. In contrast, the parietal lobe has a larger meta expression range of -1.5 to 1.35, with an interquartile range(Q1-Q3) of -0.53 to 0.29. Therefore, as gene expression distribution varies across brain regions, a sensible cut-off (if one was to be used) for highly and lowly expressed genes may lie at the 1st and 3rd quartiles respectively, with quartiles calculated per tissue.

The non-AD disorder microarray datasets

Nine non-AD studies were identified and acquired, of which four studies consisted of samples generated from multiple disorders and brain regions. Separating the studies by disease and tissue equated to 21 datasets consisting of 8 SCZ, 6 BD, 4 HD, 2 MDD and 1 PD dataset with a total of 1166 samples after QC. The demographics of the non-AD datasets is provided in *Table 3*, and a complete list of DEG's is provided in Supplementary Table 2. SCZ and BD were the only disorders with expression data available across all four brain regions, and the frontal lobe brain region was the only region with expression data available from all non-AD disorders identified in this study.

Summary of non-AD brain disorder meta-analyses DEG counts

A second meta-analysis was performed on all non-AD disorders, and similarly to the AD meta-analysis, datasets were grouped into the TL, FL, PL and CB brain regions. An overview of the non-AD meta-analysis results are provided in *Table 4*.

The meta-analysis expression profiles

As described in the methods, three primary expression signatures were derived from the meta-analyses for each of the four brain regions: - 1) "AD expression profile", 2) "AD-specific expression profile" and 3) "common neurological disorder expression profile". The numbers of significant DEG's in each of the three expression signatures are provided in *Table 5.*

The DEG's from the "AD expression profile" in the TL brain region were not significantly DE in any other disorder included in this study. Hence, the "AD expression profile" and the "AD-specific expression profile" contained the same 323 genes for the TL brain region. The "AD-specific expression profile" for all four brain regions is provided in *Supplementary Table 3.*

The "common neurological disorder expression profile" within the four brain regions consisted of very little or no DEG's (except for the parietal lobe); hence, the downstream analysis did not yield any statistically significant results of biological relevance. We find little robust

Table 1: Characteristics of individual AD studies processed in this meta-analysis

Data repository	Accession details (Publication)	Microarray platform	BeadArray	Tissue source (as stated in the original study publication)	Meta-Analysis brain region mapping	Number of samples after QC	
						AD (M/F)	Control (M/F)
ArrayExpress	E-GEOD-118553	Illumina	HumanHT-12 v4	Entorhinal Cortex	Temporal Lobe	35 (14/21)	21 (12/9)
				Cerebellum	Cerebellum	38 (10/28)	19 (5/14)
				Frontal Cortex	Frontal Lobe	38 (13/25)	22 (11/11)
				Temporal Cortex	Temporal Lobe	51 (21/30)	29 (21/8)
ArrayExpress	E-GEOD-48350 ([41])	Affymetrix	Human Genome U133 Plus 2.0	Entorhinal Cortex	Temporal Lobe	11 (6/5)	38 (21/17)
				Hippocampus	Temporal Lobe	15 (8/7)	41 (22/19)
				Postcentral Gyrus	Parietal Lobe	19 (11/8)	33 (20/13)
				Superior Frontal Gyrus	Frontal Lobe	17 (8/9)	38 (22/16)
ArrayExpress	E-GEOD-29378 ([7])	Illumina	HumanHT-12 v3	Hippocampus CA1	Temporal Lobe	16 (9/7)	16 (11/5)
				Hippocampus CA3	Temporal Lobe	15 (9/6)	16 (11/5)
ArrayExpress	E-GEOD-36980 ([8])	Affymetrix	Human Gene 1.0 ST	Frontal Cortex	Frontal Lobe	14 (7/7)	17 (9/8)
				Hippocampus	Temporal Lobe	7 (3/4)	10 (5/5)
				Temporal Cortex	Temporal Lobe	10 (5/5)	19 (8/11)
ArrayExpress	E-GEOD-28146 ([19])	Affymetrix	Human Genome U133 Plus 2.0	Hippocampus CA1	Temporal Lobe	15 (4/11)	8 (5/3)
ArrayExpress	E-GEOD-1297 ([42])	Affymetrix	Human Genome U133A	Hippocampus	Temporal Lobe	19 (4/11)	9 (6/3)
ArrayExpress	E-GEOD-5281 ([21])	Affymetrix	Human Genome U133 Plus 2.0	Entorhinal Cortex	Temporal Lobe	10 (4/6)	13 (11/2)
				Hippocampus CA1	Temporal Lobe	10 (4/6)	13 (10/3)
				Medial Temporal Gyrus	Temporal Lobe	16 (10/6)	12 (8/4)
				Posterior Cingulate	Parietal Lobe	9 (4/5)	13 (10/3)
				Superior Frontal Gyrus	Frontal Lobe	23 (13/10)	11 (7/4)
AMP	syn3157225 ([43])	Illumina	Whole-Genome DASL HT	Temporal Cortex	Temporal Lobe	189 (93/96)	186 (116/70)
				Cerebellum	Cerebellum	169 (87/82)	171 (113/58)
AMP	syn4552659 ([44])	Affymetrix	Human Genome U133A	Frontal Pole	Frontal Lobe	25 (6/19)	7 (4/3)
				Precentral Gyrus	Frontal Lobe	20 (5/15)	3 (1/2)

				Inferior Frontal Gyrus	Frontal Lobe	19 (5/14)	4 (1/3)
				Dorsolateral Prefrontal Cortex	Frontal Lobe	19 (4/15)	8 (4/4)
				Superior Parietal Lobule	Parietal Lobe	11 (2/9)	5 (2/3)
				Prefrontal Cortex	Frontal Lobe	23 (7/16)	4 (2/2)
				Parahippocampal Gyrus	Temporal Lobe	18 (5/13)	7 (3/4)
				Hippocampus	Temporal Lobe	20 (5/15)	5 (2/3)
				Inferior Temporal Gyrus	Temporal Lobe	20 (5/15)	6 (3/3)
				Middle Temporal Gyrus	Temporal Lobe	15 (4/11)	7 (4/3)
				Superior Temporal Gyrus	Temporal Lobe	15 (3/12)	8 (4/4)
				Temporal Pole	Temporal Lobe	25 (7/18)	6 (3/3)
AMP	syn4552659 ([44])	Affymetrix	Human Genome U133B	Frontal Pole	Frontal Lobe	26 (8/18)	7 (4/3)
				Precentral Gyrus	Frontal Lobe	18 (4/14)	3 (1/2)
				Inferior Frontal Gyrus	Frontal Lobe	21 (5/16)	5 (2/3)
				Dorsolateral Prefrontal Cortex	Frontal Lobe	20 (5/15)	8 (4/4)
				Superior Parietal Lobule	Parietal Lobe	16 (5/11)	5 (3/2)
				Prefrontal Cortex	Frontal Lobe	23 (7/16)	4 (2/2)
				Parahippocampal Gyrus	Temporal Lobe	19 (7/12)	7 (3/4)
				Hippocampus	Temporal Lobe	22 (6/16)	5 (2/3)
				Inferior Temporal Gyrus	Temporal Lobe	21 (6/15)	7 (4/3)
				Middle Temporal Gyrus	Temporal Lobe	23 (8/15)	7 (4/3)
				Superior Temporal Gyrus	Temporal Lobe	23 (4/19)	8 (4/4)
				Frontal Pole	Frontal Lobe	26 (8/18)	7 (4/3)

Nine publicly available AD studies were identified and acquired for this study. Separating the studies by tissue resulted in 46 datasets, each containing AD and healthy control samples. The brain tissue in each of the 46 datasets was mapped to their corresponding cerebral cortex (temporal lobe, frontal lobe or parietal lobe) or the cerebellum. Due to limited phenotypic information in publicly available data, the reported gender was predicted from gene expression if clinical gender was unavailable. Abbreviations: M = Male, F=Female.

Table 2: Summary of AD study meta-analysis DEG's

Brain region	Number of datasets	Number of samples (case/control)	AW.OC Significant DEGs (FDR adjusted p ≤ 0.05)
Temporal lobe	14	850 (419/431)	323
Frontal lobe	4	180 (92/88)	460
Parietal lobe	2	74 (28/46)	1736
Cerebellum	2	397 (207/190)	867

Twenty-two AD datasets containing a total of 1501 samples remained in this study after QC. The case/control numbers represent the total number of AD/healthy controls subjects across all datasets within a particular brain region. The number of significant genes was identified through a combining p-value method known as Adaptively Weighted with One-sided Correction (AW.OC).

evidence of shared biology based on this data analysis and therefore, exclude all results generated from the "common neurological disorder expression profile" from this paper; however, we provide the complete list of significantly DEG's within this profile in *Supplementary Table 4.*

Common differentially expressed genes across multiple brain regions in AD

AD is known to affect all brain regions through the course of the disease, although not to the same degree, similar transcriptomic changes across all brain regions were deemed disease-specific, while perturbations in a single brain region were considered to be tissue-specific. We were particularly interested in disease-specific transcriptomic changes and therefore decided to focus on genes that were found to be consistently DE across multiple brain regions.

Meta-analysis of the AD datasets identified a total of 2495 unique genes as significantly DE. The distribution of these genes across the four brain regions is shown in *Figure 1.* Forty-two genes were found to be perturbed across all four brain regions and can be grouped into three sets (*Figure 2*). The first group (Gene set 1) are expressed consistently in the same direction across all four brain regions and can be regarded as disease-specific. The second group (Gene set 2) are expressed in the same direction in the TL, FL and PL, but expression is reversed in the CB brain region, a region suggested to be spared from AD pathology [4,5]. This expression pattern suggests these genes may be involved in AD pathology. Finally, the third group (Gene set 3) are inconsistently expressed across the four brain regions are most likely tissue-specific or even false-positives.

From the forty-two genes significantly differentially expressed across all brain regions, seven genes were DE in the same direction and belong to the "AD-specific expression profile", that is, these seven genes (down-regulated NDUFS5, SOD1, SPCS1 and up-regulated OGT, PURA, RERE, ZFP36L1) were consistently perturbed in all AD brain regions and not in any other brain region of any other disorder used in this study and can be considered unique to AD brains. The expression of these seven genes across AD brains is shown in *Figure 3.*

Differentially expressed genes in brain regions affected by AD histopathology

In AD, the TL, FL and PL are known to be affected by both plaques and tangles, while the CB brain region is rarely reported to be affected. In addition to identifying genes DE across all brain regions and reversed in the CB brain region, we were also interested in genes perturbed in the TL, FL and PL and not the CB. These genes may also play a role in general AD histopathology and could be new therapeutic targets in preventing or curing AD.

Fifty-five genes were found to be significantly DE in TL, FL and PL but not the CB, of which sixteen were expressed in the same direction and were not DE in the other brain disorders used in this study. Ten of these genes (ALDOA,

Table 3: Characteristics of individual non-AD studies included in this meta-analysis

Data repository	ArrayExpress Accession details (Publication)	Microarray Platform	BeadArray	Disorder	Sample source (as stated in the original study publication)	Mapping to brain region	Number of samples after QC	
							AD (M/F)	Control (M/F)
ArrayExpress	E-GEOD-12649 ([43])	Affymetrix	Human Genome U133A	Bipolar Disorder	Prefrontal Cortex	Frontal Lobe	33 (16/17)	34 (25/9)
				Schizophrenia	Prefrontal Cortex	Frontal Lobe	33 (25/8)	32 (24/8)
ArrayExpress	E-GEOD-17612 ([44])	Affymetrix	Human Genome U133 Plus 2.0	Schizophrenia	Prefrontal Cortex	Frontal Lobe	27 (18/9)	22 (11/11)
ArrayExpress	E-GEOD-20168 ([45])	Affymetrix	Human Genome U133A	Parkinson's Disease	Prefrontal Cortex	Frontal Lobe	14 (7/7)	16 (11/5)
ArrayExpress	E-GEOD-21138 ([46])	Affymetrix	Human Genome U133 Plus 2.0	Schizophrenia	Prefrontal Cortex	Frontal Lobe	25 (21/4)	28 (23/5)
ArrayExpress	E-GEOD-21935 ([47])	Affymetrix	Human Genome U133 Plus 2.0	Schizophrenia	Temporal Cortex	Temporal Lobe	22 (12/10)	19 (10/9)
ArrayExpress	E-GEOD-35978 ([48])	Affymetrix	Human Gene 1.0 ST	Bipolar Disorder	Cerebellum	Cerebellum	32 (16/16)	46 (29/17)
				Schizophrenia	Cerebellum	Cerebellum	43 (31/12)	46 (29/17)
				Bipolar Disorder	Parietal Lobe	Parietal Lobe	40 (24/16)	45 (32/13)
				Schizophrenia	Parietal Lobe	Parietal Lobe	51 (37/14)	36 (26/10)
ArrayExpress	E-GEOD-3790 ([49])	Affymetrix	Human Genome U133A	Huntingdon's Disease	Frontal Lobe	Frontal Lobe	36 (22/14)	27 (19/8)
				Huntingdon's Disease	Cerebellum	Cerebellum	38 (22/16)	27 (16/11)
			Human Genome U133B	Huntingdon's Disease	Cerebellum	Cerebellum	38 (23/15)	27 (16/11)
				Huntingdon's Disease	Frontal Lobe	Frontal Lobe	37 (21/16)	29 (19/10)
ArrayExpress	E-GEOD-5388 ([50])	Affymetrix	Human Genome U133A	Bipolar Disorder	Prefrontal Cortex	Frontal Lobe	30 (16/14)	29 (23/6)
ArrayExpress	E-GEOD-53987 ([51])	Affymetrix	Human Genome U133 Plus 2.0	Bipolar Disorder	Prefrontal Cortex	Frontal Lobe	17 (10/7)	19 (11/8)
				Major Depressive Disorder	Prefrontal Cortex	Frontal Lobe	16 (9/7)	18 (10/8)
				Schizophrenia	Prefrontal Cortex	Frontal Lobe	14 (7/7)	19 (11/8)
				Bipolar Disorder	Hippocampus	Temporal Lobe	18 (11/7)	17 (9/8)
				Major Depressive Disorder	Hippocampus	Temporal Lobe	16 (9/7)	17 (9/8)
				Schizophrenia	Hippocampus	Temporal Lobe	15 (9/6)	18 (10/8)

Nine publicly available non-AD studies were identified and acquired. Separating the studies by tissue resulted in 21 datasets. Each dataset contained both diseased and complimentary healthy controls. The brain tissue in each of the 21 datasets was mapped to their corresponding cerebral cortex (temporal lobe, frontal lobe or parietal lobe) or the cerebellum. Due to limited phenotypic information in publicly available data, the reported gender was predicted from gene expression if clinical gender was unavailable. Abbreviations: M = Male, F=Female.

Table 4: Summary of non-AD study meta-analysis DEG's

Brain region	Number of BD datasets (case/control)	Number of Schizophrenia datasets (case/control)	Number of HD datasets (case/control)	Number of MDD datasets (case/control)	Number of PD datasets (case/control)	Total number of datasets (case/control)	AW.OC Significant DEGs (FDR adjusted p ≤ 0.05)
Temporal lobe	1 (18/17)	2 (37/37)	0	1 (16/17)	0	4 (71/71)	51
Frontal lobe	3 (80/82)	4 (99/101)	2 (73/56)	1 (16/18)	1 (14/16)	11 (282/273)	149
Parietal lobe	1 (40/45)	1 (51/36)	0	0	0	2 (91/81)	2611
Cerebellum	1 (32/46)	1 (43/46)	2 (76/54)	0	0	4 (151/146)	177

The table illustrates the non-AD dataset and sample distribution across the four brain regions. Disease abbreviations are as follows: BD=Bipolar Disease, HD= Huntington's Disease, MDD=Major Depressive Disorder and PD=Parkinson's Disease. The case/control numbers represent the total number of diseased and healthy control subjects within a disease group and brain region. For instance, "3 (80/82)" for BD datasets in the Frontal lobe region indicates three BD datasets with a combined total of 80 BD and 82 complimentary healthy control subjects. The number of significant DEG's was identified through a combining p-value method known as Adaptively Weighted with One-sided Correction (AW.OC).

Table 5: Summary of DEGs in each expression signature and brain region

Expression Profile	Cerebellum	Frontal lobe	Parietal lobe	Temporal lobe	Total (unique)
AD	867	460	1736	323	2494
Non-AD	177	149	2611	51	2809
AD-specific	828	435	1023	323	1994
Common	39	25	713	0	755
Total (unique)	1005	584	3642	374	-

The "AD" expression profile represents genes identified as DE in the AD vs control meta-analysis. The "non-AD" expression profile represents genes identified as DE in the non-AD meta-analysis. The "AD-specific" expression profile is a list of genes DE in AD and no other disorder, and the "common" expression profile is a list of genes DE in all mental disorder used in this study. Each expression profile is brain region-specific. The "Total (unique)" represents a unique list of the total number of genes identified as significantly DE across brain regions or expression profiles

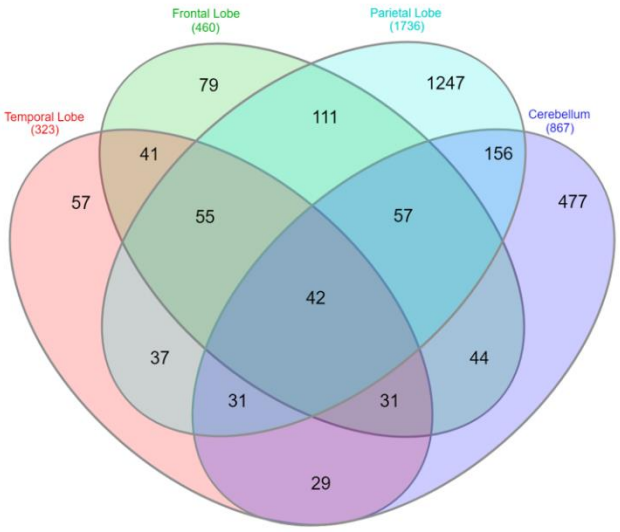


Figure 1: Overlap of DEG's in the AD expression profile across brain regions. Forty-two genes were observed to be significantly differentially expressed across all four AD brain regions.

GABBR1, TUBA1A, GAPDH, DNM3, KLC1, COX6C, ACTG1, CLTA, SLC25A5) were consistently down-regulated, and six genes (PRNP, FDFT1, RHOQ, B2M, SPP1, WAC) were consistently up-regulated in AD.

Furthermore, from the forty-two genes identified as significantly DE across all four AD brain regions, ten genes were in consensus in their expression across the TL, FL and PL brain region but expression is reversed in the cerebellum. Only 3 of these genes (UBA1, EIF4H and CLDND1) belong to the "AD-specific expression profile", and all three genes were significantly down-regulated in the TL, FL and PL, but significantly up-regulated in the CB brain region (see Gene set 2 in Figure 2).

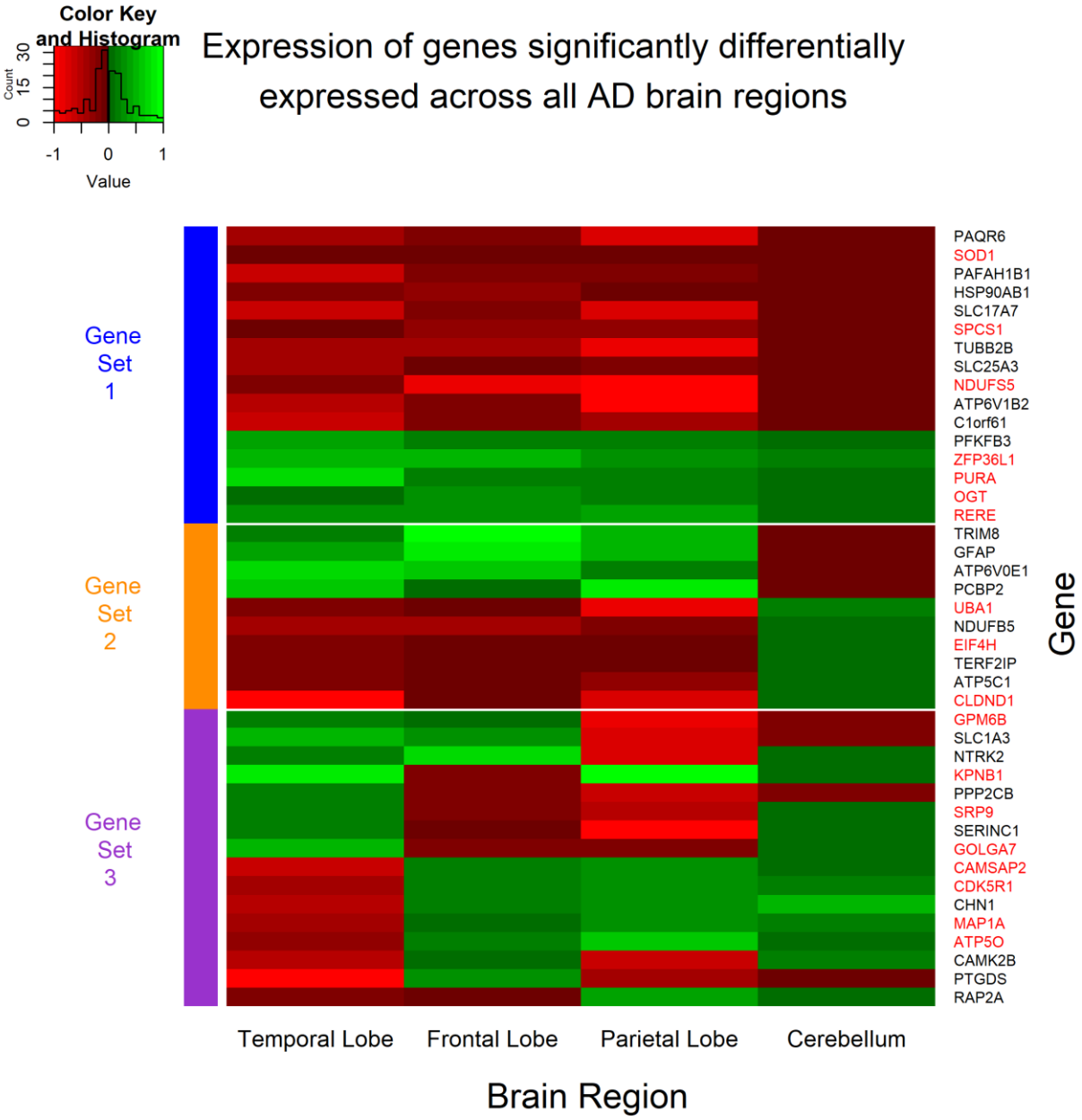


Figure 2: Expression pattern of genes significantly (FDR adjusted p-value ≤0.05) differentially expressed across all four AD brain regions. The expression values for each gene was obtained from the meta-summary calculations. Red cells represent down-regulated genes, and green cells represent up-regulated genes. Forty-two genes were observed to be significantly perturbed across all four AD brain regions and can be grouped into three "sets". Gene set 1 represents genes which are perturbed consistently in the same direction across all AD brain regions and can be considered disease-specific. Gene set 2 represents genes consistent in expression in the TL, FL and PL brain regions, but reversed in the CB brain region; a region often referred to be free from AD pathology. Finally, Gene set 3 represents genes which are significant DE across all four brain regions; however, directional change is not consistent across the brain regions and may represent tissue-specific genes or even false positive. The gene names highlighted in red are genes perturbed in AD and not in any other disorder used in this study and are deemed "AD-specific".

Microarray gene expression profiling in RNA-Seq data

The 7 genes (**NDUFS5**, **SOD1**, **SPCS1**, **OGT**, **PURA**, **RERE**, **ZFP36L1)** consistently expressed across all brain regions and the 19 genes (**ALDOA, GABBR1, TUBA1A, GAPDH, DNM3, KLC1, COX6C, ACTG1, CLTA, SLC25A5**, **PRNP, FDFT1, RHOQ, B2M, SPP1, WAC, UBA1**, **EIF4H**, **CLDND1**) consistently expressed in the TL, FL and PL and not in the CB or reversed in the CB, were queried in the web-based platform Agora to compare RNA-Seq based expression profiling. The results are provided in *Table 6*. Agora failed to provide expression profiling for 17/26 genes, however, from the data available, the genes observed to be consistently expressed across all brain regions based on

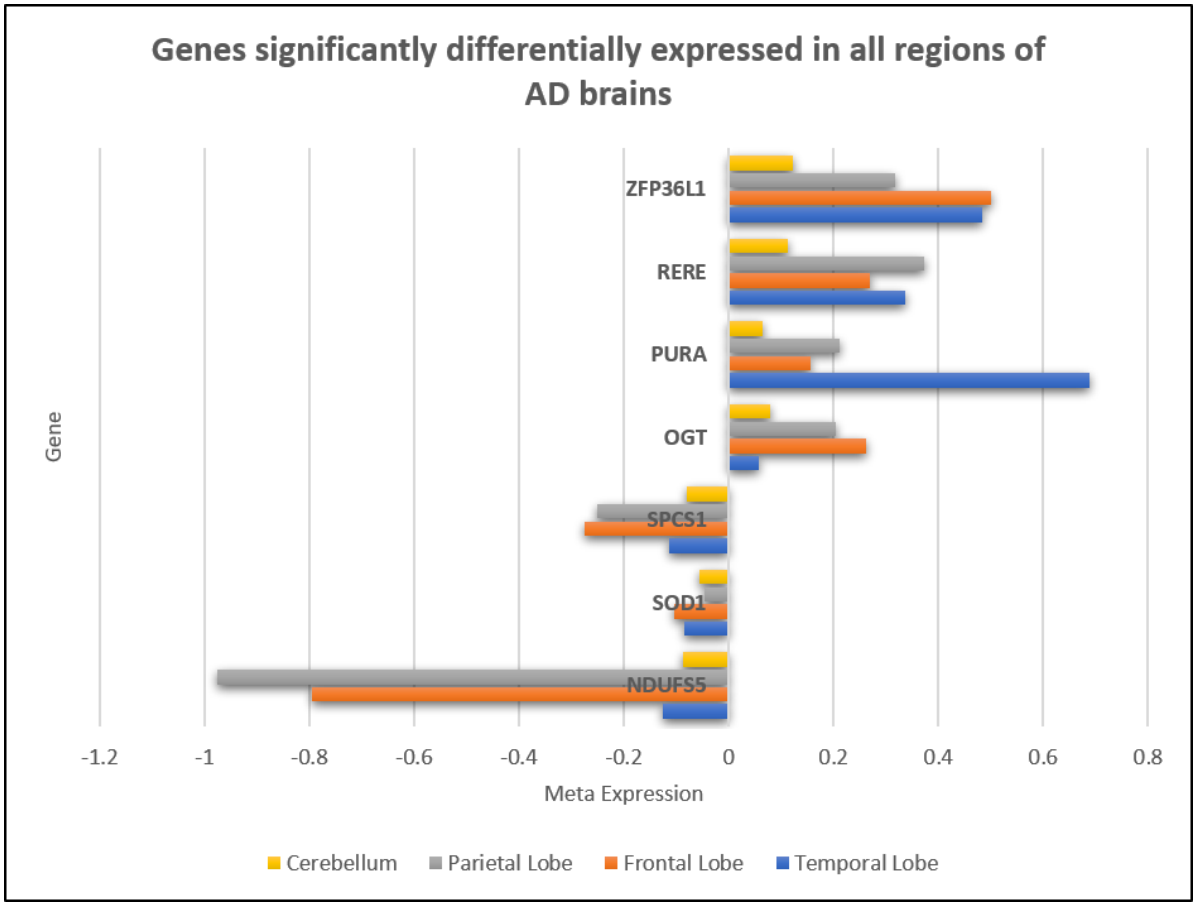


Figure 3: Seven genes consistently significantly differentially expressed in the same direction in all regions of AD brains but not in Schizophrenia, Bipolar Disorder, Huntington's disease, Major Depressive Disorder or Parkinson's disease brains. These seven genes can be assumed to be unique to AD brains and may play an important role in disease mechanisms.

microarray data are relatively mirrored in RNA-Seq data, specifically genes **SPCS1**, **PURA** and **ZFP36L1.** RNA-Seq data was available for only 6/19 genes (**DNM3, COX6C, ACTG1, CLTA, RHOQ** and **B2M**) expressed in brain regions affected by hallmark AD pathology (TL, FL and PL), and were all relatively consistent in directional change across AD brain regions, including the CB, a characteristic undesired by genes which may be associated with hallmark AD pathology.

"AD Expression Profile" functional gene set enrichment and GO analysis

Gene set enrichment analysis of the "AD expression profile" identified 205, 197, 98, and 45 biological pathways significantly enriched in the TL, FL, PL and CB brain regions respectively (*Supplementary Table 5*). There were ten pathways significantly enriched in all four brain regions, of which eight are involved in the "**metabolism of protein**" (specifically the translation process, the most significant being in CB brain region with a q-value=1.11e-7), one involved in "**adenosine ribonucleotides de novo biosynthesis**" (TL q-value = 0.007, FL q-value = 7.56e-5, PL q-value = 0.04, CB q-value = 0.03) and one involved in the "**digestive system**" (TL q-value = 0.02, FL q-value = 0.02, PL q-value = 0.01, CB p-value = 0.02).

When excluding the CB brain region, 42 pathways were significantly enriched in the remaining three brain regions, of which five pathways obtained an FDR adjusted significance p-value of ≤ 0.01. The five pathways are "**Alzheimer's disease**" (TL q-value 6.53e-4, FL q-value = 0.02, PL q-value = 0.01),

Table 6: Microarray gene expression compared to RNA-Seq gene expression

Gene	Microarray				RNA-Seq						
	TL	FL	PL	CB	TL			FL			CB
					TCX	STG	PHG	IFG	FP	DLPFC	
NDUFS5	-0.127	-0.796	-0.975	-0.087							
SOD1	-0.085	-0.104	-0.047	-0.055							
SPCS1	-0.113	-0.276	-0.251	-0.079	-0.192		-0.177			-0.124	-0.317
OGT	0.057	0.262	0.205	0.079							
PURA	0.688	0.157	0.211	0.065						0.103	0.318
RERE	0.337	0.270	0.373	0.114							
ZFP36L1	0.484	0.502	0.317	0.123	0.603	0.389	0.619				0.355
ALDOA	-1.389	-0.090	-0.091								
GABBR1	-0.864	-0.116	-0.162								
TUBA1A	-0.728	-0.286	-0.191								
GAPDH	-0.719	-0.095	-0.759								
DNM3	-0.611	-0.154	-0.781		-0.361		-0.119				0.141
KLC1	-0.530	-0.115	-0.056								
COX6C	-0.504	-0.360	-0.651			-0.140	-0.200				-0.220
ACTG1	-0.467	-0.165	-0.326							-0.250	
CLTA	-0.153	-0.097	-0.183		-0.091						-0.112
SLC25A5	-0.147	-0.122	-0.262								
PRNP	0.115	0.168	0.130								
FDFT1	0.158	0.196	0.283								
RHOQ	0.227	0.128	0.330		0.465		0.310			0.190	0.029
B2M	0.556	0.216	0.212		0.430						0.258
SPP1	0.935	0.557	0.316								
WAC	1.111	0.257	0.154								
UBA1	-0.213	-0.090	-0.827	0.134							
EIF4H	-0.201	-0.062	-0.093	0.063							
CLDND1	-1.100	-0.113	-0.734	0.086							

The 7 genes consistently expressed across all brain regions and the 19 genes consistently expressed in the TL, FL and PL and not/reversed in the CB were queried in the web-based platform Agora to compare RNA-Seq expression. Only significantly (p-value ≤ 0.05) DE genes after multiple correction are shown. Red cells represent down-regulated genes in AD, green cells represent up-regulated genes in AD, white cells represent genes not significantly DE, and grey cells are when data is not available. Abbreviations are as follows; TL = Temporal Lobe, FL = Frontal Lobe, PL = Parietal Lobe, CB = Cerebellum, TCX = Temporal Cortex, STG = Superior Temporal Gyrus, PHG = Parahippocampal Gyrus, IFG = Inferior frontal gyrus, FP = Frontal pole, DLPFC = Dorsolateral prefrontal cortex.

"Electron Transport Chain" (TL q-value = 0.006, FL q-value = 2.95e-5, PL q-value = 3.69e-5), "Oxidative phosphorylation" (TL q-value =1.77e-4, FL q-value = 4.99e-8, PL q-value = 4.18e-05), "Parkinson's disease" (TL q-value =8.57e-4, FL q-value = 1.59e-6, PL q-value = 1.77e-6) and "Synaptic vesicle cycle" (TL q-value = 5.19e-4, FL q-value = 3.82e-7, PL q-value = 2.03e-4).

The biological GO analysis identified 384, 417, 216, and 72 biological components as significantly enriched in the TL, FL, PL and CB brain region respectively (*Supplementary Table 6*). There were 36 pathways significantly enriched across all four brain regions at a p-value threshold of ≤ 0.05 and nine at an FDR adjusted significant p-value threshold of ≤ 0.01. These nine processes are "cellular component biogenesis" (TL q-value =1.38e-4, FL q-value = 0.002, PL q-value = 5.86e-4, CB q- value = 0.006), "cellular component organization" (TL q-value = 1.96e-8, FL q-value 0.004), "interspecies interaction between organisms" (TL q-value = 1.13e-41.85e-4, FL q-value = 8.73e-5, PL q-value = 5.59e-5 , CB q-value = 0.002), "multi-organism cellular process" (TL q-value = , FL q-value = 4.72e-5, PL q-value = 8.04e-5, CB q-value = 0.002), "nervous system development" (TL q-value = 1.64e-7, FL q-value = 5.90e-14, PL q-value = 3.82e-8, CB q-value = 0.01), "organonitrogen compound metabolic process" (TL q-value = 0.002, FL q-value = 1.56e-5, PL q-value = 1.02e-5, CB q-value = 0.002), "symbiosis, encompassing, mutualism through parasitism" (TL q-value = 4.04e-4, FL q- value = 1.92e-4, PL q-value = 3.18e-4, CB q-value = 0.004), "translational initiation" (TL q-value = 0.007, FL q-value = 0.006, PL q-value = 2.41e-4, CB q-value = 5.24e-6), and "viral process" (TL q-value = 2.82e-4, FL q-value = 1.17e-4, PL q-value = 3.18e-4, CB q-value =0.002). Excluding the CB brain region resulted in 84 common biological components being significantly enriched across the remaining three brain regions.

"AD-Specific Expression Profile" functional gene set enrichment and GO analysis

Analysis of the "AD-specific expression profile" identified 205, 196, 40 and 42 pathways as significantly enriched in the TL, FL, PL and CB brain region respectively in the GSEA analysis *(Supplementary Table 7*). The analysis identified six significantly enriched pathways across all four brain regions, and all are involved in "metabolism of protein" (specifically the translation process, with the most significant pathway being in the PL brain region with a q-value = 8.92e-7). The same six pathways were identified when the CB region was excluded.

The GO analysis identified 384, 344, 36 and 72 significantly enriched biological components for the TL, FL, PL and CB brain region respectively. Only four common biological components were significantly enriched across all four brain regions, and all are indicative of interspecies interactions including viral. Excluding the CB identifies only "neural nucleus development" (TL q-value = 5.35e-5, FL q-value = 0.007, PL q-value = 0.003) as an additional component being enriched. The complete biological GO analysis results are provided in *Supplementary Table 8.*

Network analysis hub gene identification

PPI networks were generated for each expression profile and in each of the four brain regions (TL, FL, PL and CB) to identify genes whose protein product interacts with other protein products from the same expression profile. Genes with more interactions than expected are referred to as hub genes and may be of biological significance.

Temporal lobe hub genes

PPI network analysis was performed on the expression profiles of the TL brain region to identify key hub genes. The "AD expression profile" and the "AD-specific expression profile" both consisted of the same 323 DEG's which represented 282 seed proteins with 716 edges (interactions between proteins). Two significant key hub genes were identified; the down-regulated Polyubiquitin-C (UBC, p-value = 1.57e-30) and the up-regulated Small Ubiquitin-related Modifier 2 (SUMO2, p-value = 3.7e-4).

Frontal lobe hub genes

The FL "AD expression profile" consisted of 460 DEG's which represented 272 seed proteins and 620 edges. Two significant key hub genes were identified; up-regulated Amyloid Precursor Protein (APP, p-value = 1.98e-08) and down-regulated Heat Shock Protein 90-alpha (HSP90AA1, p-value = 0.003). Using the "AD-specific expression profile" identified the same two key hub genes, with APP reaching a significant p-value of 2.11e-09.

Parietal lobe hub genes

The PL "AD expression profile" consisted of 1736 DEG's which represented 1437 seed proteins and 5720 edges. Similar to the TL and FL, two significant key hub genes were identified; down-regulated Cullin-3 (CUL3, p-value = 1.84e-10) and down-regulated UBC (p-value = 1.84e-10). Using the "AD-specific expression profile" (1023 DEGs, 810 seed proteins and 2351 edges) identified UBC as the only key hub gene, with a more significant p-value of 1.84e-10. The CUL3 gene is no longer a significant key hub gene in the network.

Cerebellum hub genes

The CB "AD expression profile" consisted of 867 DEG's which represented 548 seed proteins and 1419 edges. Four significant key hub genes were identified; up-regulated APP (p-value =4.24e-26), down-regulated Ribosomal Protein 2 (RPS2, p-value = 4.24e-26), down-regulated SUMO2 (p-value = 4e-05), and up-regulated Glycyl-TRNA Synthetase (GARS, p-value = 0.0207). Using the "AD-specific expression profile" for the same brain region identified APP (p-value = 3.44e-26), RPS2 (p value= 6.61e-06), and SUMO2 (p-value = 3.78e-06) as the key hub genes only. The GARS gene is no longer a key hub gene in the network.

DISCUSSION

In this study, we acquired eighteen publicly available microarray gene expression studies covering six neurological and mental health disorders; AD, BD, HD, MDD, PD and SCZ. Data was generated on seven different expression BeadArrays and across two different microarray technologies (Affymetrix and Illumina). The eighteen studies consisted of 3984 samples extracted from 22 unique brain regions which equated to 67 unique datasets when separating by disorder and tissue. However, due to study and sample outlier analysis, only 43 datasets (22 AD, 6 BD, 4 HD, 2 MDD, 1 PD and 8 SCZ) totalling 2,667 samples passed QC. We grouped the AD datasets by tissue, into the TL, FL, PL and CB brain regions to perform the largest microarray AD meta-analysis known to date to our knowledge, which identified 323, 460, 1736 and 867 significant DEG's respectively. Furthermore, we incorporated transcriptomic information from other neurological and mental health disorders to subset the initial findings to 323, 435, 1023, and 828 significant DEG's that were specifically perturbed in the TL, FL, PL and CB brain regions respectively of AD subjects.

Genes specifically perturbed across AD brain regions

Seven genes (down-regulated NDUFS5, SOD1, SPCS1 and up-regulated OGT, PURA, RERE, ZFP36L1) were DE in AD brains and not DE in the other disorders used in this study. We deemed these seven protein-coding genes as "AD-specific". The expression patterns of three genes (SPCS1, PURA and ZFP36L1) were relatively mirrored in RNA-Seq data; however, it is important to note the RNA-Seq data does not contain expression profiling for the PL region, and it also contains three specific brain regions within the TL (temporal cortex, superior temporal gyrus, Parahippocampal gyrus) and FL (Inferior frontal gyrus, frontal pole and dorsolateral prefrontal cortex). Nevertheless, the SPCS1 gene was observed to be consistently down-regulated across all hierarchical AD brain regions available in both the microarray and RNA-Seq data. In addition, based on a network of genomics and epigenomic elements in the region of this gene, in combination with phenotypes, the AMP-AD consortia have nominated SPCS1 as a druggable target for AD treatment.

Three of the "AD-specific" genes (NDUFS5, SOD1 and OGT) have been previously associated with AD. Down-regulated NADH Dehydrogenase Ubiquinone Fe-S Protein 5 (NDUFS5) gene is part of the human mitochondrial respiratory chain complex; a process suggested to be disrupted in AD in multiple studies [45]. A study investigating blood-based AD biomarkers identified 13 genes, including NDUFS5, which was capable of predicting AD with 66% accuracy (67% sensitivity and 75% specificity) in an independent cohort of 118 AD and 118 control

subjects [46]. The perturbation in NDUFS5 expression in the blood and brains of AD subjects suggests this gene may have potential as an AD biomarker and warrants further investigation.

Down-regulated Superoxide Dismutase 1 (SOD1) gene encodes for copper and zinc ion binding proteins which contribute to the destruction of free superoxide radicals in the body and is also involved in the function of motor neurons [provided by RefSeq, Jul 2008]. Mutations in this gene have been heavily implicated as causes of familial amyotrophic lateral sclerosis (ALS) [47] and have also been associated with AD risk [48]. A recent study discovered SOD1 deficiency in an amyloid precursor protein-overexpressing mouse model accelerated Aβ oligomerisation and also caused Tau phosphorylation [49]. They also stated SOD1 isozymes were significantly decreased in human AD patients, and we can now confirm SOD1 is significantly under-expressed at the mRNA level in human AD brains as well.

The up-regulated O-Linked N-Acetyl Glucosamine Transferase (OGT) gene encodes for a glycosyltransferase that links N-acetylglucosamine to serine and threonine residues (O-GlcNAc). O-GlcNAcylation is the post-translational modification of O-GlcNAc and occurs on both neuronal tau and APP. Increased brain O-GlcNAcylation has been observed to protect against tau and amyloid-β peptide toxicity [50]. A mouse study has demonstrated a deletion of the encoding OGT gene causes an increase in tau phosphorylation [51]. In this study, we observe a significant increase in OGT gene expression throughout human AD brains, including the cerebellum where tangles are rarely reported, suggesting OGT gene is most likely not solely responsible for the formation of tangles.

OGT and O-GlcNAcase (OGA) enzymes facilitate O-GlcNAc cycling, and levels of GlcNAc have also been observed to be increased in the parietal lobe of AD brains [52]. Appropriately, OGA inhibitors have been tested for treating AD with promising preliminary results [53], prompting further investigation into targeting OGT for AD treatment.

Genes involved in AD histopathology

The CB brain region is known to be free from tau pathology and occasionally free from plaques. We exploited the CB brain region as a secondary control to identify sixteen genes (ALDOA, GABBR1, TUBA1A, GAPDH, DNM3, KLC1, COX6C, ACTG1, CLTA, SLC25A5, PRNP, FDFT1, RHOQ, B2M, SPP1, WAC) DE specifically in TL, FL and PL and not the CB brain region of AD subjects. RNA-Seq data was available for 6 of these genes (DNM3, COX6C, ACTG1, CLTA, RHOQ and B2M) and all 6 genes failed to replicate expression patterns observed with microarray data. Nevertheless, DNM3 gene has been previously associated with AD pathology based on proteomic data DNM3 gene encodes a member of a family of guanosine triphosphate (GTP)-binding proteins that associate with microtubules and are involved in vesicular transport. A proteomic study identified a module of co-expressed proteins, which included DNM3, as negatively correlated with BRAAK staging [54]. Although DNM3 gene expression based on microarray and RNA-Seq data are in disagreement in the CB brain region, a region used in this study to aid in determining whether a gene may be involved with AD pathology, an independent proteomic study demonstrated DNM3 might indeed be association with AD pathology. This suggests all 6 genes which failed replication in RNA-Seq data may still be associated with AD pathology and require further confirmation.

An additional 9 genes (GABBR1, GAPDH, PRPN, FDFT1, KLC1, TUBA1A, CLTA, COX6C, SLC25A5), where expression profiling based on RNA-Seq data was unavailable, have also been previously associated with AD, of which four genes (GABBR1, GAPDH, PRPN and FDFT1) have individually been suggested to be involved with the pathogenesis of the disease. GABBR1 gene encodes a receptor for gamma-aminobutyric acid (GABA), which is the primary inhibitory neurotransmitter in the human central nervous system. As observed in this study, the GABBR1 gene has been previously reported to be down-regulated in AD brains [16]. GABBR1 receptors

are prominent in neuronal soma, where NFT formation is known to accumulate. A study examined the immunohistochemical localisation and distribution of GABABR1 protein in the hippocampus of AD subjects and observed a negative correlation with NFT formation and suggested an increase or stable expression of GBBR1 could contribute to neuronal resistance to the disease process [55].

GAPDH gene encodes for a member of the glyceraldehyde-3-phosphate dehydrogenase protein family, which catalyses an essential step in the carbohydrate metabolism. GAPDH has been shown to interact with Aβ precursor protein but not cleaved Aβ and has been proposed to be directly involved in tau aggregation and NFT formation in AD [56–58]. The PRNP gene encodes for the prion protein, a membrane glycosylphosphatidylinositol-anchored glycoprotein that tends to aggregate into rod-like structures. Mutations in the PRNP gene has been associated with AD and prion protein has also been suggested to be involved in the pathogenesis of AD [59]. FDFT1 gene encodes a membrane-associated enzyme located at a branch point in the mevalonate pathway, which generates isoprenoids that have been found to be positively correlated with tau pathology [60]. KLC1 gene encodes for Kinesin Light Chain 1 which transports various cargos such as vesicles, mitochondria, and the Golgi complex along microtubules. An immunoblotting study observed decrease expression of kinesin light chains (KLCs) in the frontal cortex of AD subjects but not in the cerebellum of the same subjects [61]. TUBA1A gene encodes for Tublin Alpha 1a, which has been observed to be perturbed in AD [62], and CLTA gene encodes for clathrin Light Chain A, which has been observed to be perturbed in AD as well [63]. COX6C and SLC25A5 gene encodes for products which interact with mitochondria and mitochondrial dysfunction in AD has been suggested on numerous occasions [45,64,65].

We identified an additional three AD-specific genes (UBA1, EIF4H and CLDND1) which were significant DE in all four brain regions. However, the genes were down-regulated in the TL FL and PL but up-regulated in the CB brain region. Ubiquitin-Like Modifier Activating Enzyme 1 (UBA1) encodes for a protein that catalyses the first step in ubiquitin conjugation to mark cellular proteins for degradation. Eukaryotic Translation Initiation Factor 4H (EIF4H) encodes for a translation initiation factors, which functions to stimulate the initiation of protein synthesis at the level of mRNA utilisation and Claudin Domain Containing 1 (CLDND1) is a transmembrane protein of tight junctions found on endothelial cells [66]. As the cerebellum is the only brain region spared from tangle formation and occasionally from plaque, we suggest these 19 genes (ALDOA, GABBR1, TUBA1A, GAPDH, DNM3, KLC1, COX6C, ACTG1, CLTA, SLC25A5, PRNP, FDFT1, RHOQ, B2M, SPP1, WAC, UBA1, EIF4H and CLDND1)) could potentially be associated with AD histopathology.

Translation of proteins perturbed specifically in AD brains

Functional gene set enrichment analysis of the "AD expression profile" revealed more pathways were significantly perturbed in the TL, followed by the FL, PL and CB, which is the general route AD pathology is known to spread through the brain. We originally observed ten biological pathways being enriched across all AD brain regions, which included biological pathways likely to be irrelevant such as the "digestive system". However, when incorporating transcriptomic information from non-AD disorders, we were able to refine the AD expression signature to specific genes perturbed in AD only. This resulted in the enrichment of pathways only involved in the "metabolism of proteins", specifically the translation process which has been previously suggested in be associated with AD on numerous occasions [10,11,14–17] We now suggest this may be a biological process specifically disrupted in AD brains, and not BD, HD, MDD, PD or SCZ brains.

Previous biological perturbations observed in AD are only associated with the temporal lobe brain region.

Previous AD studies have consistently suggested the immune response [10–13]

protein transcription/translation regulation [10,11,14–17], calcium signalling [10,18,19], MAPK signalling [7,16] chemical synapse [7,18,19], neurotransmitter [11,18,19] various metabolism pathways [11,16,17,20–23] are disrupted in AD. We observe the same pathways enriched in our meta-analysis; however, only in the TL brain region, a brain region often heavily investigated in AD. Except for "**metabolism of proteins**", we did not observe any of these pathways significantly enriched across all of the four brain regions, suggesting these pathways observed to be perturbed in previous studies may be tissue-specific rather than disease-specific.

Interspecies interactions possibly involved in AD

Gene Ontology analysis on the "AD expression profile" identified nine different biological components enriched across all four brain regions. However, when we remove genes perturbed in other neurological or mental health disorders, we only observe four biological components as significantly enriched, and all four were indicative of interspecies interactions. AD brains have a prominent inflammatory component which is characteristic of infection, and many microbes have been implicated in AD, notably herpes simplex virus type 1 (HSV1), Chlamydia pneumonia, and several types of spirochaete [67]. A very recent study also identified common viral species in normal and ageing brains, with an increased human herpesvirus 6A and human herpesvirus 7 in AD brains [68]. Furthermore, Aβ has been suggested to be an antimicrobial peptide and has been shown to protect against fungal and bacterial infections [69]. Thus, the accumulation of Aβ may be part of the brains defence mechanism against infections. Although a controversial theory, we also observe a viral component in AD brains, and as a result of this meta-analysis, further suggest this maybe AD-specific and warrants further investigation.

Network analysis identifies AD-specific APP UBC and SUMO2 hub genes

Network analysis identified five (**APP, HSP90AA1, UBC, SUMO2** and **RPS2**) significant hub genes specific to AD brain regions. **APP, UBC** and **SUMO2** gene appear as hub genes in multiple brain regions. The **APP** gene encodes for a cell surface receptor transmembrane amyloid precursor protein (APP) that is cleaved by secretases to form a number of peptides. Some of these peptides are secreted and can bind to the acetyltransferase complex APBB1/TIP60 to promote transcriptional activation, while others form the protein basis of the amyloid plaques in AD brains. In addition, two of the peptides are antimicrobial peptides, having been shown to have bactericidal and antifungal activities [provided by RefSeq, Aug 2014]. Changes in APP functions have been suggested to play an essential role in the lack of AB clearance, ultimately leading to the formation of plaques [70].

UBC (ubiquitin-C) gene encodes for a Polyubiquitin-C protein which is part of the ubiquitin-proteasome system (UPS), the primary intracellular protein quality control system in eukaryotic cells. UPS has an immense impact on the amyloidogenic pathway of APP processing that generates Abeta [71]. A recent GWAS study identified **UBC** as a novel LOAD gene, and through network analysis also identified **UBC** as a key hub gene. The study validated their findings in a **UBC** C. elegans model to discover **UBC** knockout accelerated age-related AB toxicity [72]. We also observe the **UBC** gene being down-regulated and as a key hub gene in multiple regions of human AD brains, further providing evidence of its key role in AD.

Small Ubiquitin-Like Modifier 2 (**SUMO2**) gene encodes for a protein that binds to target proteins as part of a post-translational modification system, a process referred to as SUMOylation [73]. However, unlike ubiquitin, which targets proteins for degradation, this protein is involved in a variety of cellular processes, such as nuclear transport, transcriptional regulation, apoptosis, and protein stability [provided by RefSeq, Jul 2008]. Early studies have indicated that the **SUMO** system is likely altered with AD-type pathology, which may impact Aβ levels and tau

aggregation [73]. Genetic studies have supported this theory with a GWAS study linking SUMO-related genes to LOAD [74], with further studies showing that the two natively unfolded proteins, tau and α-synuclein, are sumoylated in vitro [75]. We identified SUMO2 as a significant key hub gene in both the human TL and the CB brain region. However, what makes this discovery interesting is that SUMO2 is significantly up-regulated in the TL, a region where both plaques and tangles can be observed, but significantly down-regulated in the CB, where only plaques have been occasionally observed, but tangles never reported. The up-regulation of SUMO2 gene may play a vital role in the formation of tangles, and further investigation into this gene is warranted.

Limitations

Although this study presents novel insights to AD-specific transcriptomic changes in the human brain, limitations to this study must be addressed. Firstly, we meta-analysed a total of 22 AD and 21 non-AD datasets, and many of these datasets lacked necessary experimental processing or basic phenotypic information such as technical batches, RNA integrity numbers (RIN), age, NFT's, clinical gender, or ethnicity, all of which can have confounding effects. To address this, we incorporated recommended best practices to estimate and correct for both known and hidden batch effects using SVA and COMBAT to ensure data is comparable between experiments and studies. However, this does not guarantee that all technical variation is completely removed.

Secondly, the terminology used to label brain tissue varied across studies, with some reporting a broad region such as the "hippocampus" used in study E-GEOD-48350, while others were particular to the tissue layer, such as "hippocampus CA3" in study E-GEOD-29378. We, therefore, decided to map all brain tissue as mentioned in each dataset publication to their hierarchical cerebral cortex lobe (TL, FL and PL) and the CB. The mapping procedure was completed using publicly available literature defined knowledge, and we assume tissues within these brain regions are relatively comparable to infer AD-associated histopathological changes.

Thirdly, this study relied on publicly available transcriptomic data, and as previous research has heavily investigated brain regions known to be at the forefront of disease manifestation, this led to unbalanced datasets per brain region in both the AD and non-AD meta-analysis. Subsequently, the AD meta-analysis consisted of 14, 4, 2, and 2 datasets for the TL, FL, PL and CB brain regions respectively, with the PL brain region consisting of only 74 samples (28 AD and 46 controls) in total. In addition, the non-AD meta-analysis lacked expression signatures form all non-AD diseases across all brain regions (except for FL). Nevertheless, the brain regions most affected by each disorder was captured in this study, suggesting we most likely were able to capture key brain transcriptomic changes relating to each disorder. Furthermore, as AD is known to affect all brain regions, albeit not to the same extent, we focus on transcriptomic changes observed across all brain regions that are also not observed in any brain region of the non-AD subjects, ensuring we capture transcriptomic signatures unique to AD brains.

Fourthly, the advances in next sequencing technologies (RNA-Seq) which are capable of profiling the whole transcriptome, thus not limited by the pre-defined probes based on known sequencing, would be ideal for disease discoveries. However, AD and mental health studies profiled through RNA-Seq is somewhat limited in the public domain, and those that have published DE results are based on small sample numbers, which would fail our selection criteria, such as [76–79]. In addition, these studies lack the same brain regions and mental health disorders covered in this meta-analysis. Nevertheless, we were able to query our genes of interest in the largest known AD RNA-Seq web-based database (Agora) which contains DE results from over 2100 human brain samples; however, expression profiling was unavailable for 17/26 genes, and DE on the parietal lobe was unavailable. Therefore, this study was unable to validate all findings in RNA-Seq data and additional experimental validation, such as

RT-PCR is still required to support this study's results.

Finally, we assume the non-AD datasets are comparable through meta-analysis, and by identifying common expression signatures that are not associated with individual disease mechanisms may represent false positives or even a general signature for "brain disorder". Removing this signature from the AD meta-analysis expression profile may result in transcriptomic changes specific to AD brains, revealing more relevant changes to the underlying disease mechanism rather than general diseases. Under this assumption, we observe more relevant and refined biological enrichment results. For example, we originally observed ten biological pathways enriched across all AD brain regions, including biological pathways such as the "digestive system". However, by refining the AD expression signature by removing genes perturbed in non-AD disorders, only pathways involved in the "metabolism of proteins" remain, which has been previously suggested in be associated with AD on numerous occasions [10,11,14–17] This observation provides strong evidence of our assumption of incorporating non-AD diseases in this study to infer AD-specific changes as valid.

Conclusion

We present the most extensive human AD brain microarray transcriptomic meta-analysis study to date, incorporating, brain regions both affected and partially spared by AD pathology, and utilise related non-AD disorders to infer AD-specific brain changes. This led to the identification of seven genes specifically perturbed across all AD brain regions and are considered disease-specific, nineteen genes specifically perturbed in AD brains which could play a role in AD neuropathology, and the refinement of GSEA and GO analysis results to identify specific biological pathways and components specific to AD. These AD-specific changes may provide new insights into the disease mechanisms, thus making a significant contribution towards understanding the disease.

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: http://dx.doi.org/ 10.3233/JAD-181085.

Acknowledgements

This study presents independent research supported by the NIHR BioResource Centre Maudsley at South London and Maudsley NHS Foundation Trust (SLaM) & Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, NIHR, Department of Health or King's College London.

RJBD and SJN are supported by 1. Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome Trust. 2. The National Institute for Health Research University College London Hospitals Biomedical Research Centre.

References

[1] Prince M, Albanese Emiliano, Prina Matthew (2014) World Alzheimer Report 2014 Dementia and Risk Reduction. Alzheimer's Dis. Int.

[2] Rajan KB, Wilson RS, Weuve J, Barnes LL, Evans DA (2015) Cognitive impairment 18 years before clinical diagnosis of Alzheimer disease dementia. Neurology 85, 898–904.

[3] Serrano-Pozo A, Frosch MP, Masliah E, Hyman BT (2011) Neuropathological alterations in Alzheimer disease. Cold Spring Harb. Perspect. Med. 1, 1–23.

[4] Convit a., De Asis J, De Leon MJ, Tarshish CY, De Santi S, Rusinek H (2000)

Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease. Neurobiol. Aging 21, 19–26.

[5] Jacobs HIL, Hopkins DA, Mayrhofer HC, Bruner E, van Leeuwen FW, Raaijmakers W, Schmahmann JD (2017) The cerebellum in Alzheimer's disease: evaluating its role in cognitive decline. Brain.

[6] Zhang M, Yao C, Guo Z, Zou J, Zhang L, Xiao H, Wang D, Yang D, Gong X, Zhu J, Li Y, Li X (2008) Apparently low reproducibility of true differential expression discoveries in microarray studies. Bioinformatics 24, 2057–2063.

[7] Miller JA, Woltjer RL, Goodenbour JM, Horvath S, Geschwind DH (2013) Genes and pathways underlying regional and cell type changes in Alzheimer's disease. Genome Med. 5, 48.

[8] Hokama M, Oka S, Leon J, Ninomiya T, Honda H, Sasaki K, Iwaki T, Ohara T, Sasaki T, LaFerla FM, Kiyohara Y, Nakabeppu Y (2014) Altered expression of diabetes-related genes in Alzheimer's disease brains: the Hisayama study. Cereb. Cortex 24, 2476–88.

[9] Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P (2005) Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. Nucleic Acids Res. 33, 5914–5923.

[10] Sekar S, McDonald J, Cuyugan L, Aldrich J, Kurdoglu A, Adkins J, Serrano G, Beach TG, Craig DW, Valla J, Reiman EM, Liang WS (2015) Alzheimer's disease is associated with altered expression of genes involved in immune response and mitochondrial processes in astrocytes. Neurobiol. Aging 36, 583–591.

[11] Li Y, Wu Z, Jin Y, Wu A, Cao M, Sun K, Jia X, Chen M (2012) Analysis of hippocampal gene expression profile of Alzheimer 's disease model rats using genome chip bioinformatics. 7, 332–340.

[12] Lambert JC, Grenier-Boley B, Chouraki V, Heath S, Zelenika D, Fievet N, Hannequin D, Pasquier F, Hanon O, Brice A, Epelbaum J, Berr C, Dartigues JF, Tzourio C, Campion D, Lathrop M, Amouyel P (2010) Implication of the immune system in Alzheimer's disease: evidence from genome-wide pathway analysis. J. Alzheimer's Dis. 20, 1107–1118.

[13] Chen J, Xie C, Zhao Y, Li Z, Xu P (2016) Gene expression analysis reveals the dysregulation of immune and metabolic pathways in Alzheimer ' s disease. 7, 1–6.

[14] Liu T, Ren D, Zhu X, Yin Z, Jin G, Zhao Z, Robinson D, Li X, Wong K, Cui K, Zhao H, Wong STC (2013) Transcriptional signaling pathways inversely regulated in Alzheimer's disease and glioblastoma multiform. Sci. Rep. 3, 3467.

[15] Li X, Long J, He T, Belshaw R, Scott J (2015) Integrated genomic approaches identify major pathways and upstream regulators in late onset Alzheimer's disease. Sci. Rep. 5, 12393.

[16] Puthiyedth N, Riveros C, Berretta R, Moscato P (2016) Identification of differentially expressed genes through integrated study of Alzheimer's disease affected brain regions. PLoS One 11, 1–29.

[17] Godoy JA, Rios JA, Zolezzi JM, Braidy N, Inestrosa NC (2014) Signaling pathway cross talk in Alzheimer's disease. Cell Commun. Signal. 12, 23.

[18] Ramanan VK, Kim S, Holohan K, Shen L, Nho K, Risacher SL, Foroud TM, Mukherjee S, Crane PK, Aisen S, Petersen RC, Weiner MW, Saykin AJ (2013) Genome-wide pathway analysis of memory impairment in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort implicates gene candidates, canonical pathways, and networks. 6, 634–648.

[19] Blalock EM, Buechel HM, Popovic J, Geddes JW, Landfield PW (2011) Microarray

analyses of laser-captured hippocampus reveal distinct gray and white matter signatures associated with incipient Alzheimer's disease. J. Chem. Neuroanat. 42, 118–126.

[20] Paolo G Di, Kim T (2012) Linking Lipids to Alzheimer ' s Disease : Cholesterol and Beyond. Aging (Albany. NY). 12, 284–296.

[21] Liang WS, Reiman EM, Valla J, Dunckley T, Beach TG, Grover A, Niedzielko TL, Schneider LE, Mastroeni D, Caselli R, Kukull W, Morris JC, Hulette CM, Schmechel D, Rogers J, Stephan DA (2008) Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. Proc. Natl. Acad. Sci. U. S. A. 105, 4441–6.

[22] Ishii K, Sasaki M, Kitagaki H, Yamaji S, Sakamoto S, Matsuda K, Mori E (1997) Reduction of cerebellar glucose metabolism in advanced Alzheimer's disease. J. Nucl. Med. 38, 925–928.

[23] Oshiro S, Morioka MS, Kikuchi M (2011) Dysregulation of iron metabolism in Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis. Adv. Pharmacol. Sci. 2011,.

[24] Edwards YJK, Beecham GW, Scott WK, Khuri S, Bademci G, Tekin D, Martin ER, Jiang Z, Mash DC, ffrench-Mullen J, Pericak-Vance MA, Tsinoremas N, Vance JM (2011) Identifying consensus disease pathways in Parkinson's disease using an integrative systems biology approach. PLoS One 6,.

[25] Chandrasekaran S, Bonchev D (2013) A network view on Parkinson's disease. Comput. Struct. Biotechnol. J. 7, e201304004.

[26] Clelland CL, Read LL, Panek LJ, Nadrich RH, Bancroft C, Clelland JD (2013) Utilization of Never-Medicated Bipolar Disorder Patients towards Development and Validation of a Peripheral Biomarker Profile. PLoS One 8, 1–11.

[27] ohn I. Nurnberger Jr, MD P, Daniel L. Koller P, Jeesun Jung P, J., HowardEDenberg P, Tatiana Foroud P, Ilaria Guella P, Marquis P. Vawter P, And, John, Kelsoe R (2015) Identification of Pathways for Bipolar Disorder A Meta-analysis. Curr. Drug Targets 16, 700–710.

[28] Chen H, Wang N, Zhao X, Ross CA, O'Shea KS, Mcinnis MG (2013) Gene expression alterations in bipolar disorder postmortem brains. Bipolar Disord. 15, 177–187.

[29] Khanzada N, Butler M, Manzardo A (2017) GeneAnalytics Pathway Analysis and Genetic Overlap among Autism Spectrum Disorder, Bipolar Disorder and Schizophrenia. Int. J. Mol. Sci. 18, 527.

[30] William R. Markesbery (2010) Neuropathologic Alterations in Mild Cognitive Impairment: A Review. J Alzheimers Dis 19, 221–228.

[31] Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, Weiss-Solís DY, Duque R, Bersini H, Nowé A (2013) Batch effect removal methods for microarray gene expression data integration: A survey. Brief. Bioinform. 14, 469–490.

[32] Hackstadt AJ, Hess AM (2009) Filtering for increased power for microarray data analysis. BMC Bioinformatics 10, 11.

[33] Oldham MC, Langfelder P, Horvath S (2012) Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. BMC Syst. Biol. 6, 63.

[34] Kang DD, Sibille E, Kaminski N, Tseng GC (2012) MetaQC: Objective quality control and inclusion/exclusion criteria for genomic meta-analysis. Nucleic Acids Res. 40, 1–14.

[35] Chang L-C, Lin H-M, Sibille E, Tseng GC (2013) Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. BMC Bioinformatics 14, 368.

[36] Dalman MR, Deeter A, Nimishakavi G, Duan Z-H (2012) Fold change and p-value cutoffs significantly alter microarray interpretations.

file:///home/sbagew/PhD/workspace/publications/Co-expressionNetworks/Afroza/Fold_change/Gene_Category/GSE19677/ddq018.pdf 13, S11.

[37] St G, Shtokalo D, Tackett MR, Yang Z, Vyatkin Y, Milos PM, Seilheimer B, Mccaffrey TA, Kapranov P (2013) On the importance of small changes in RNA expression. Methods 63, 18–24.

[38] Kamburov A, Wierling C, Lehrach H, Herwig R (2009) ConsensusPathDB - A database for integrating human functional interaction networks. Nucleic Acids Res. 37, 623–628.

[39] Xia J, Benner MJ, Hancock REW (2014) NetworkAnalyst - Integrative approaches for protein-protein interaction network analysis and visual exploration. Nucleic Acids Res. 42, 167–174.

[40] Zaki N, Mora A (2015) A comparative analysis of computational approaches and algorithms for protein subcomplex identification. Sci. Rep. 4, 4262.

[41] Bonilla E, Tanji K, Hirano M, Vu TH, Dimauro S, Schon EA (1999) Mitochondrial involvement in Alzheimer's disease. Biochim. Biophys. Acta - Bioenerg. 1410, 171–182.

[42] Voyle N, Keohane A, Newhouse S, Lunnon K, Johnston C, Soininen H, Kloszewska I, Mecocci P, Tsolaki M, Vellas B, Lovestonea S, Hodges A, Kiddle S, Dobson RJB (2016) A pathway based classification method for analyzing gene expression for Alzheimer's disease diagnosis. J. Alzheimer's Dis. 49, 659–669.

[43] Rosen DR, Siddique T, Patterson D, Figlewicz DA, Sapp P, Hentati A, Donaldson D, Goto J, O'Regan JP, Deng HX, Daniel R. Rosen, Teepu Siddique DP, Rosen DR, Siddique T, Patterson D, Figlewicz DA, Sapp P, Hentati A, Donaldson D, Goto J, O'Regan JP, Deng HX, Daniel R. Rosen, Teepu Siddique DP (1993) Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. Nature 362, 59–62.

[44] Spisak K, Klimkowicz-Mrowiec A, Pera J, Dziedzic T, Golenia A, Slowik A (2015) rs2070424 of the SOD1 gene is associated with risk of alzheimer's disease. Neurol. Neurochir. Pol. 48, 342–345.

[45] Murakami K, Murata N, Noda Y, Tahara S, Kaneko T, Kinoshita N, Hatsuta H, Murayama S, Barnham KJ, Irie K, Shirasawa T, Mice SAD, Early S, Loss M (2011) SOD1 (Copper / Zinc Superoxide Dismutase) Deficiency Drives Amyloid Protein Oligomerization and Memory Loss in Mouse Model of Alzheimer Disease. 286, 44557–44568.

[46] Zhu Y, Shan X, Yuzwa SA, Vocadlo DJ (2014) The emerging link between O-GlcNAc and Alzheimer disease. J. Biol. Chem. 289, 34472–34481.

[47] O'Donnell N, Zachara NE, Hart GW, Marth JD (2004) Protein Glycosylation Is a Requisite Modification in Somatic Cell Function and Embryo Viability Ogt -Dependent X-Chromosome-Linked Protein Glycosylation Is a Requisite Modification in Somatic Cell Function and Embryo Viability. Mol Cell Biol 24, 1680–1690.

[48] Förster S, Welleford AS, Triplett JC, Sultana R, Schmitz B, Butterfield DA (2014) Increased O-GlcNAc levels correlate with decreased O-GlcNAcase levels in Alzheimer disease brain. Biochim. Biophys. Acta - Mol. Basis Dis. 1842, 1333–1339.

[49] Yuzwa SA, Shan X, Jones BA, Zhao G, Woodward ML, Li X, Zhu Y, McEachern EJ, Silverman MA, Watson N V., Gong CX, Vocadlo DJ (2014) Pharmacological inhibition of O-GlcNAcase (OGA) prevents cognitive decline and amyloid plaque formation in bigenic tau/APP mutant mice. Mol. Neurodegener. 9, 42.

[50] Seyfried NT, Dammer EB, Swarup V, Nandakumar D, Duong DM, Yin L, Deng Q, Nguyen T, Hales CM, Wingo T, Glass J, Gearing M, Thambisetty M, Troncoso JC, Geschwind DH, Lah JJ, Levey AI (2017) A Multi-network

Approach Identifies Protein-Specific Co-expression in Asymptomatic and Symptomatic Alzheimer's Disease. Cell Syst. 4, 60–72.e4.

[51] Iwakiri M, Mizukami K, Ikonomovic MD, Ishikawa M, Hidaka S, Abrahamson EE, DeKosky ST, Asada T (2005) Changes in hippocampal GABABR1 subunit expression in Alzheimer's patients: association with Braak staging. Acta Neuropathol. 109, 467–474.

[52] El Kadmiri N, Slassi I, El Moutawakil B, Nadifi S, Tadevosyan A, Hachem A, Soukri A (2014) Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and Alzheimer's disease. Pathol. Biol. 62, 333–336.

[53] Butterfield DA, Hardas SS, Lange MLB (2010) Oxidatively modified glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and Alzheimer's disease: many pathways to neurodegeneration. J. Alzheimers. Dis. 20, 369–93.

[54] Itakura M, Nakajima H, Kubo T, Semi Y, Kume S, Higashida S, Kaneshige A, Kuwamura M, Harada N, Kita A, Azuma Y-T, Yamaji R, Inui T, Takeuchi T (2015) Glyceraldehyde-3-phosphate Dehydrogenase Aggregates Accelerate Amyloid-β Amyloidogenesis in Alzheimer Disease. J. Biol. Chem. 290, 26072–87.

[55] Zhang W, Jiao B, Xiao T, Pan C, Liu X, Zhou L, Tang B, Shen L (2016) Mutational analysis of PRNP in Alzheimer's disease and frontotemporal dementia in China. Sci. Rep. 6, 38435.

[56] Pelleieux S, Picard C, Lamarre-Théroux L, Dea D, Leduc V, Tsantrizos YS, Poirier J (2018) Isoprenoids and tau pathology in sporadic Alzheimer's disease. Neurobiol. Aging 65, 132–139.

[57] Morel M, Héraud C, Nicaise C, Suain V, Brion J-P (2012) Levels of kinesin light chain and dynein intermediate chain are reduced in the frontal cortex in Alzheimer's disease: implications for axoplasmic transport. Acta Neuropathol. 123, 71–84.

[58] Lachen-Montes M, Zelaya MV, Segura V, Fernández-Irigoyen J, Santamaría E (2017) Progressive modulation of the human olfactory bulb transcriptome during Alzheimer´s disease evolution: novel insights into the olfactory signaling across proteinopathies. Oncotarget 8, 69663–69679.

[59] Nakamura Y, Takeda M, Yoshimi K, Hattori H, Hariguchi S, Kitajima S, Hashimoto S, Nishimura T (1994) Involvement of clathrin light chains in the pathology of Alzheimer's disease. Acta Neuropathol. 87, 23–31.

[60] Swerdlow RH, Khan SM (2004) A "mitochondrial cascade hypothesis" for sporadic Alzheimer's disease. Med. Hypotheses 63, 8–20.

[61] Iwamoto K, Bundo M, Kato T (2005) Altered expression of mitochondria-related genes in postmortem brains of patients with bipolar disorder or schizophrenia, as revealed by large-scale DNA microarray analysis. Hum. Mol. Genet. 14, 241–253.

[62] Krause G, Winkler L, Mueller SL, Haseloff RF, Piontek J, Blasig IE (2008) Structure and function of claudins. 1778, 631–645.

[63] Sethi AJ, Wikramanayake RM, Angerer RC, Range RC, Angerer LM (2016) Microbes and Alzheimer's Disease. 335, 590–593.

[64] Readhead B, Haure-Mirande JV, Funk CC, Richards MA, Shannon P, Haroutunian V, Sano M, Liang WS, Beckmann ND, Price ND, Reiman EM, Schadt EE, Ehrlich ME, Gandy S, Dudley JT (2018) Multiscale Analysis of Independent Alzheimer's Cohorts Finds Disruption of Molecular, Genetic, and Clinical Networks by Human Herpesvirus. Neuron 99, 64–82.e7.

[65] Kumar DK V., Choi SH, Washicosky KJ, Eimer WA, Tucker S, Ghofrani J, Lefkowitz A, McColl G, Goldstein LE, Tanzi RE, Moir RD (2016) Amyloid- peptide protects against microbial infection in mouse and worm models of Alzheimers disease. Sci. Transl. Med. 8, 340ra72-340ra72.

[66] O'Brien RJ, Wong PC (2011) Amyloid Precursor Protein Processing and Alzheimer's Disease. Annu. Rev. Neurosci. 34, 185–204.

[67] Hong L, Huang H-C, Jiang Z-F (2014) Relationship between amyloid-beta and the ubiquitin–proteasome system in Alzheimer's disease. Neurol. Res. 36, 276–282.

[68] Mukherjee S, Russell JC, Carr DT, Burgess JD, Younkin MA, Serie DJ, Boehme KL, Kauwe JSK, Naj AC, Fardo DW, Dickson DW, Montine TJ, Ertekin-Taner N, Kaeberlein MR, Crane PK, Allen M, Serie DJ, Boehme KL, Kauwe JSK, Naj AC, Fardo DW, Dickson DW, Montine TJ, Ertekin-Taner N, Kaeberlein MR, Crane PK (2017) Systems biology approach to late-onset Alzheimer's disease genome-wide association study identifies novel candidate genes validated using brain expression data and Caenorhabditis elegans experiments. Alzheimer's Dement. 1–10.

[69] Lee L, Sakurai M, Matsuzaki S, Arancio O, Fraser P (2013) SUMO and alzheimer's disease. NeuroMolecular Med. 15, 720–736.

[70] Grupe A, Abraham R, Li Y, Rowland C, Hollingworth P, Morgan A, Jehu L, Segurado R, Stone D, Schadt E, Karnoub M, Nowotny P, Tacey K, Catanese J, Sninsky J, Brayne C, Rubinsztein D, Gill M, Lawlor B, Lovestone S, Holmans P, O'Donovan M, Morris JC, Thal L, Goate A, Owen MJ, Williams J (2007) Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. Hum. Mol. Genet. 16, 865–873.

[71] Dorval V, Fraser PE (2006) Small ubiquitin-like modifier (SUMO) modification of natively unfolded proteins tau and α-synuclein. J. Biol. Chem. 281, 9919–9924.

[72] Annese A, Manzari C, Lionetti C, Picardi E, Horner DS, Chiara M, Caratozzolo MF, Tullo A, Fosso B, Pesole G, Erchia AMD, D'Erchia AM, Erchia AMD (2018) Whole transcriptome profiling of Late-Onset Alzheimer's Disease patients provides insights into the molecular changes involved in the disease. Sci. Rep. 8, 1–15.

[73] Magistri M, Velmeshev D, Makhmutova M, Faghihi MA (2015) Transcriptomics Profiling of Alzheimer's Disease Reveal Neurovascular Defects, Altered Amyloid-β Homeostasis, and Deregulated Expression of Long Noncoding RNAs. J. Alzheimer's Dis. 48, 647–665.

[74] Bennett JP, Keeney PM (2018) RNA-Sequencing Reveals Similarities and Differences in Gene Expression in Vulnerable Brain Tissues of Alzheimer's and Parkinson's Diseases. J. Alzheimer's Dis. reports 2, 129–137.

[75] Mills JD, Nalpathamkalam T, Jacobs HIL, Janitz C, Merico D, Hu P, Janitz M (2013) RNA-Seq analysis of the parietal cortex in Alzheimer's disease reveals alternatively spliced isoforms related to lipid metabolism. Neurosci. Lett. 536, 90–95.

[76] Berchtold NC, Coleman PD, Cribbs DH, Rogers J, Gillen DL, Cotman CW (2013) Synaptic genes are extensively downregulated across multiple brain regions in normal human aging and Alzheimer's disease. Neurobiol. Aging 34, 1653–1661.

[77] Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Landfield PW (2004) Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc. Natl. Acad. Sci. 101, 2173–2178.

[78] Zou F, Chai HS, Younkin CS, Allen M, Crook J, Pankratz VS, Carrasquillo MM, Rowley CN, Nair AA, Middha S, Maharjan S, Nguyen T, Ma L, Malphrus KG, Palusak R, Lincoln S, Bisceglio G, Georgescu C, Kouri N, Kolbert CP, Jen J, Haines JL, Mayeux R, Pericak-Vance MA, Farrer LA, Schellenberg GD, Petersen RC, Graff-Radford NR, Dickson DW, Younkin SG, Ertekin-Taner N (2012) Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. PLoS Genet. 8,.

[79] Wang M, Roussos P, McKenzie A, Zhou X, Kajiwara Y, Brennand KJ, De Luca GC, Crary JF, Casaccia P, Buxbaum JD, Ehrlich M, Gandy S, Goate A, Katsel P, Schadt E, Haroutunian V, Zhang B (2016) Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. Genome Med. 8, 1–21.

3.1 Additional limitations: risks of biases

Chapter 3: Analysis of Alzheimer's Disease brain transcriptomic data, is an article which was published in the Journal of Alzheimer's Disease. The following is a discussion of the risks and biases associated with the study that was not included in the original publication. Bias in research is the systematic error or deviation from the truth in data collection, data analysis and result interpretation that leads to a false conclusion (Šimundić, 2013). This study combined novel and existing AD and non-AD (PD, BD, Schizophrenia, MDD and HD) brain microarray gene expression studies to infer AD specific changes. The complexity of this study leaves it open to the introduction of unintentional biases.

The first step in this study was data collection. The data used in this study was already expression profiled, with no direct input into the population selection or gene expression profiling process. The AD, PD, BD, Schizophrenia and HD samples are assumed to be a good representation of their respective population, however, without the appropriate information on data collection and associated phenotypic information, it cannot be ruled out these samples are a misrepresentation of their respective population, which would inadvertently introduce biases into the data collection and interpretation. For instance, sample collection, admission, and misclassification biases can be introduced if AD patients were diagnosed across different studies using different cognitive tests or varying test cut-off points (Storandt and Morris, 2010). Furthermore, without detailed phenotypic information on diagnosis date, it is unknown if survivor bias exists in the data. For instance, AD samples were only retained if BRAAK staging was ≥ 3, however, it is unknown how long the subjects were diagnosed with AD prior to death. Therefore, newly diagnosed AD subjects at BRAAK stage 6 may be analysed with long term AD subjects who are slow progressors or stable AD subjects with only BRAAK stage 3 pathology. This would inadvertently introduce disease heterogeneity into the mix.

Dataset selection criteria were in place to ensure only comparable datasets were acquired, reducing dataset biases. These criteria were (1) data must be generated from a neurodegenerative or mental health disorder, which would restrict to diseases that affect the brain. (2) be sampled from human brain tissue, restricting to tissues known to be affected by the disease. (3) have gene expression measured on either the Affymetrix or Illumina microarray platform, restricting to platforms that are comparable (Barnes *et al.*, 2005; Chen *et al.*, 2007; Consortium, 2008; Maouche *et al.*, 2008). (4) contain both diseased and suitably matched healthy controls in the same experimental batch, avoiding the introduction of batch effects (Chen *et al.*, 2011; Taminau *et al.*, 2012; Lazar *et al.*, 2013) and (5) contain at least 10 samples from both the diseased and control group, eliminating small underpowered datasets.

Another source of potential bias is the tissue source itself. This study relied on publicly available data, with many studies heavily investigating brain regions known to be at the forefront of disease manifestation, leading to unbalanced datasets per brain region in both the AD and non-AD disorders. This led to the AD meta-analysis consisting of 14, 4, 2, and 2 datasets for the TL, FL, PL and CB brain regions respectively. In addition, the non-AD meta-analysis lacked expression signatures across all brain regions (except FL) in the non-AD disorders. This introduces a tissue source bias where different diseases affect different tissues, hence, comparing similar tissue sources for different diseases may not reflect disease pathologies. However, the brain regions most affected by each disorder was captured in this study, and as AD is known to affect all brain regions, albeit not to the same extent, the focus on transcriptomic changes observed across all brain regions that are also not observed in any single brain region of the non-AD subjects, potentially addresses the tissue source bias.

Data analysis can be a common source of bias. For instance, when numerous methods are applied to the same data until a preferred result and/or statistically significant result is obtained (Šimundić, 2013). To avoid this type of bias, this study performed a literature review before data acquisition, to identify the most appropriate data processing and meta-analysis approaches to apply to the data. The methods applied in this study are discussed in section 1.8. Confounding variables are also a common source of biases during data analysis. Known variation in the data such as demographics (age and gender) can be accounted for during data analysis, however, the effect of unmeasured or unknown confounders can only be controlled by true study randomization (Pannucci and Wilkins, 2010). Due to the nature of the data, detailed information on each study including demographics and technical variations were somewhat limited. Therefore, to address unwanted variation, SVA and COMBAT were incorporated into the data processing procedure to estimate and eliminate variation not associated with diagnosis and gender. However, this does not guarantee that all technical variation is completely removed, and confounding biases may still remain in the data.

To avoid data interpretation biases, a fixed significant threshold (adjusted p-value ≤ 0.05) was used throughout this study to identify significant results. Genes identified as significantly and specifically perturbed in AD were further investigated in an independent dataset that was sequenced with superior sequencing technologies (RNA-Seq). This validation process adds a layer of certainty that results identified as significant by both technologies may not be by chance.

There are numerous sources of bias that can be introduced in research. This study took the relevant steps in order to reduce or eliminate the possibility of biases; nevertheless, unintentional biases may still exist, particularly within the data collection process, where no input or detailed phenotypic information on the samples exists. It is reassuring to find some results replicating in an independent AD

dataset sequenced using next-generation technologies; however, like many research output, further validation is still warranted.

Chapter 4: Automated drug-repositioning from a disease expression signature

4.1 Background

To date, no disease-modifying treatment is clinically available for AD. However, symptomatic relief is available in the form of three cholinesterase antagonists (donepezil, rivastigmine and galantamine), and an N-methyl-D-aspartate (NMDA) inhibitor known as memantine. These drugs only provide temporary symptomatic relief, do not modify the underlying disease pathology, and do not have the same effect in all patients (Calciano *et al.*, 2010; Anand, Gill and Mahdi, 2014).

4.1.1 AD drug discovery

There is a high failure rate for AD drug discovery, with more than 99% of clinical trials showing no difference between the compound of interest and the placebo, while disease-modifying therapies have a 100% failure rate (Cummings, Feldman and Scheltens, 2019). Drugs require specific characteristics to be viable as scalable treatment option for diseases, such as its toxicity, absorption, distribution, metabolism and excretion. However, for AD, an additional property is required which enables the compound to penetrate the BBB. This can complicate preclinical trials as the human BBB and its properties may not be present in animal AD model's, and therefore a compounds effect in animal models may not be reflected in human clinical trials (Cummings, Feldman and Scheltens, 2019).

One of the major challenges in AD drug discovery has been to identify the right druggable target. The current successful treatments for AD have all targeted the cholinergic system. However, since the underlying cause of the disease is currently unknown, it's been extremely difficult to develop treatments to reverse or even halt the neuropathological process of AD. The central causal hypothesis for AD involves the amyloid and tau cascade, which naturally became therapeutic targets. However, drugs targeting these systems have so far failed. Identifying the right participants in AD clinical trials is an additional important process and has most likely contributed to the lack of suitable compounds identified for AD treatment. Studies have shown many AD diagnosed patients lack Aβ deposition, with one study showing 25% of subjects diagnosed with mild AD lacked evidence of Aβ plaques when investigated with amyloid PET (Sevigny *et al.*, 2016). Therefore, these patients may not respond to compounds targeting AD pathology.

The development of new compounds targeting AD remains an important health goal. The current failures in AD clinical trials have been suggested to be caused by (1) wrong drug, (2) wrong dose, (3) wrong target, (4) wrong study design, (5) wrong outcome measures, (6) wrong analytical method, (7)

wrong stage of disease (i.e., “too late” or “too early”), (8) problems with recruitment and retention, (9) unacceptable tolerability, (10) wrong conceptual model of the disease, (11) wrong patients, and/or (12) poor study conduct (Khachaturian *et al.*, 2018).

4.1.2 Drug repositioning

The primary strategy for drug development was to screen multiple molecules for a single therapeutic target. However, the ratio of successful drugs identified against the total number of drugs has declined dramatically through the years (Iorio *et al.*, 2013). The typical time for an AD drug development program takes 13 years and costs an estimated £4.4 billion (Cummings, Reiber and Kumar, 2018). Drug repositioning (also referred to as drug repurposing) can provide an alternative strategy, where Food and Drug Administration (FDA) approved drugs are repurposed for new therapeutic targets, reducing the costs and time associated with drug development, whilst making the process accessible to academic centres along with pharmaceutical and biotechnology companies. Drug repositioning has been successfully applied to a variety of diseases, including cancer, cardiovascular disease, stress incontinence, irritable bowel syndrome, obesity, erectile dysfunction, smoking cessation, psychosis, attention deficit disorder and Parkinson's disease (Corbett *et al.*, 2012).

Typically, there are three types of drug repositioning strategies; computational approaches, biological experiments, and mixed approaches (Xue *et al.*, 2018). Computational approaches can be further subcategorised into knowledge-based and signature-based drug repositioning. Knowledge-based drug repositioning uses predetermined information such as drug-target, chemical structures, adverse effects and pathways to predict unknown mechanisms, targets or biomarkers for diseases (Yella *et al.*, 2018). In contrast, signature-based drug repositioning compares gene expression profiles from drugs directly to the target diseases to identify candidate compounds, which is the focus of this chapter.

Publicly available gene expression repositories such as GEO (Barrett *et al.*, 2013) and ArrayExpress (Brazma *et al.*, 2003) provide a wealth of disease expression profiles, while the Connectivity Map (CMap) (Lamb *et al.*, 2006) and Library of Integrated Network-based Cellular Signatures (LINCS) (Subramanian *et al.*, 2017) databases provide the most extensive publicly available drug-induced gene expression profiles. The CMap and LINCS databases contain transcriptomic profiles of dozens of cell lines treated with thousands of chemical compounds and can serve as “reference” databases. These databases provide a computational opportunity for gene-expression-based drug repositioning for many disorders, including AD.

4.1.2.1 The Connectivity Map

The CMap was first introduced in 2006 (Lamb *et al.*, 2006), with the idea to become a web-based reference database (http://www.broad.mit.edu/cmap/) containing drug-specific gene expression

profiling which can be queried using a disease-specific gene expression signature. The initial database consisted of 164 small molecules (referred to as compounds in this thesis) and was named CMap build 01. The updated build 02 consists of 1309 compounds. All experiments were performed on the Affymetrix microarray platform using the HT_HG_U133A (22,283 probesets), HG_U133A (22,277 probesets) or HT_HG_U133A_EA (21,607 probesets) GeneChips. The experiments were performed on five different cell lines which included; MCF7, ssMCF7, HL60, PC3 and SKMEL5. The MCF7 cell line is human breast epithelial adenocarcinoma, ssMCF7 is MCF7 cell lines cultured in phenol red-free DMEM, HL60 is a human promyelocytic cell line, PC3 is epithelial cell line from human prostate adenocarcinoma, and SKMEL5 is a human malignant melanoma cell line. The experiments were performed multiple times (batches), occasionally using the same compound at different concentrations, resulting in over 6000 instances (compound expression profiles). Each batch consisted of at least one calibrator compound that was referred to as the "vehicle control". Following standardised pre-processing of the raw data using MAS 5.0, gene fold change in each instance was calculated using the "vehicle control" as a comparison, genes were then ordered by decreasing fold change and assigned a rank. Thus, for a single instance expression profiled on the HT_HG_U133A GeneChip, the gene ranked one will be the most up-regulated gene while the gene ranked 22,283 would be the most down-regulated gene.

A disease-gene signature is calculated using the standard differential expression analysis, with the DEG's split into two lists; one containing up-regulated genes while the second contains down-regulated genes. Only the top 500 genes in each list can be queried in the current CMap database using a rank-ordered Kolmogorov–Smirnov (KS) test to calculate a connectivity score, which is further normalised using random permutations. The refined connectivity score ranges from -1 to +1 where positive connectivity results indicate the corresponding instance induced a similar expression profile to the disease, while a negative connectivity score indicates the corresponding instance repressed the expression profile of the disease. The degree of positive/negative connectivity scores also reflects how strongly related the disease and drug has been calculated according to the KS algorithm, while the permutation test assigns a statistical significance (Musa *et al.*, 2018).

Several studies have used the CMap to identify candidate compounds to treat several disorders (Musa *et al.*, 2018), including AD (Williams, 2012; F. Chen *et al.*, 2013; Fowler *et al.*, 2015; Siavelis *et al.*, 2016b), with a few experimentally validating the candidate compound to show association with the target disease. For instance, a study queried a gastric cancer gene expression signature in the CMap, resulting in histone deacetylase inhibitors vorinostat and trichostatin A being identified as candidate compounds. The authors experimentally validated vorinostat in vitro using cancer cell lines to demonstrate the compound significantly inhibited cell viability in a dose-dependent manner (Lim, Lim and Cho, to date,

2014). However, to date, no literature has been found to demonstrate drug repositioning using the CMap has resulted in candidate compounds being clinically used to treat diseases.

4.1.2.1.1 The CMap Limitations

The CMap has many limitations, with a significant drawback being the processing of the raw data. The CMap data were initially processed using MAS 5.0 background correction, which removes the Mismatch (MM) probe intensity from the multiple Perfect-Match (PM) intensity for each transcript to calculate an expression value for each probe independently. However, this has been reported to increase variation at low signal strengths, which has been suggested to be a result of non-specific binding of the MM probes (Irizarry *et al.*, 2003). Alternative methods, such as RMA can result in more accurate expression values and have been suggested to be superior to MAS5.0 during DE analysis (Jiang *et al.*, 2008), which is the foundation of drug repositioning the CMap. In addition, the use of a single "vehicle control" within each instance batch cannot account for technical and biological variation, which can be addressed by grouping replicated instances and performing differential expression analysis against the average change observed from a group of controls.

Furthermore, the CMap restricts users to query only 500 genes which can only consist of genes perturbed in the same direction, i.e. be either up or down-regulated. The current method of querying two separate gene lists will identify two separate lists of candidate compounds and will not take into consideration all the genes further perturbed by a drug or disease. For instance, querying a disease signature where gene A is up-regulated, and gene B is down-regulated in the current CMap database, may suggest candidate compound X which down-regulates the gene A. However, the current method does not take into consideration if compound X also further down-regulates gene B, which may further exacerbate the disease. Therefore, it is vital to consider all genes perturbed in a disease when identifying suitable candidate compounds, thus, using a single disease expression signature to query the CMap may be more effective in identifying relevant candidate compounds.

4.1.2.2 The Library of Integrated Network-based Cellular Signatures

The Library of Integrated Network-based Cellular Signatures (LINCS) database expands on the CMap database by providing approximately one million gene expression profiles generated from 22,412 different perturbations applied to 56 different cellular contexts, including 5806 genetic perturbations such as single-gene knockdowns or overexpression (Duan *et al.*, 2014). The LINCS data was generated using the L1000 technology, which is a high-throughput method to estimate genome-wide mRNA expression by measuring only 1000 genes and imputing the remaining 22,000 genes based on expression patterns observed in GEO. This makes the LINCS approach a cost-effective method for generating gene expression profiles on a large scale.

4.1.2.2.1 The LINCS Limitations

The database relies on technology which directly measures the expression of a relatively small number of genes, which is then imputed approximately 22-fold using both internal and external information. Gene expression reproducibility has always been questionable, with studies reporting low reproducibility at gene level across studies with identical samples (Zhang *et al.*, 2008). Relying on low reproducible expression data to infer expression of unknown genes can inevitably lead to poor imputation. A systematic review on the LINCS data observed 80% correlation between the L1000 and Affymetrix platform generated data prior to LINCS data imputation, however, once L1000 was imputed, the overlap of differentially expressed genes was only 30% (Musa *et al.*, 2018). The study concluded that a robust QC pipeline is required to process the LINCS data prior to addressing biological questions, and therefore, due to the issues surrounding the reliability of the LINCS database, this chapter only incorporates the CMap.

4.1.3 Alternative approaches to drug reposition the CMap

The potential capability and pitfalls of the CMap have been widely recognised, leading many studies to reprocess the CMap data and query using alternative algorithms. Some of the more common approaches are briefly discussed below.

4.1.3.1 ssCMap: Statistically significant connectivity map

Zhang et al. developed a method named statistically significant connectivity map (ssCMap) that implemented additional permutation to decrease the number of false positives. They assigned a new ranking system based on gene expression change between the compound and the control vehicle, where all genes are ranked including both up and down-regulated genes into one list. Unlike the current CMap method, this would allow the query expression signature to contain both up and down-regulated genes during the matching process. Following a connection scoring process, compounds are assigned values of -1 to +1 in relation to the input query signature, where +1 indicates similar expression profiles and -1 indicates inversely correlated expression profiles.

The method was validated using a random 25 probe gene expression signature in both the CMap and ssCMap database. The hypothesis here is that this expression signature is random and should not find any significantly relevant compounds. The CMap database identified 113 candidate compounds with positive connectivity scores and 83 with negative connectivity scores. In contrast, the ssCMap database identified no significant candidate compounds, suggesting the ssCMap reduces false positives (Zhang and Gant, 2008). Further validation for sensitivity was performed using an independent gene expression signature of HDAC inhibitors, estrogens and immunosuppressive drugs, with results indicating ssCMap has overall better sensitivity when compared to the CMap approach (Zhang and Gant, 2008).

4.1.3.2 CMapBatch: A meta-analysis of drug response

Fortney et al. adapted an established meta-analysis framework to analyse multiple gene expression signatures of disease in parallel. The method, named CMapBatch, attempts to address the heterogeneity issue of gene expression data where different gene signatures from the same disease often show poor reproducibility across studies (Fortney *et al.*, 2015). The method applies multiple gene expression signatures obtained from independent studies to the web-based CMap portal, resulting in multiple lists of candidate compounds which are combined and ranked based on how highly the compound was ranked in individual results. The authors validated their framework using 21 previously published lung cancer studies, which identified 247 candidate compounds. The authors used independent growth inhibitor data to demonstrated that compounds identified through CMapBatch were significantly better at inhibiting growth than other CMap compounds.

4.1.3.3 ProbCMap: Probabilistic drug connectivity mapping

Pakkinen and Kaski focused on common and distinct gene expression changes of the same compound on multiple different cell lines to create a probabilistic connectivity map (ProbCMap) (Parkkinen and Kaski, 2014). The ProbCMap is based on correlations of compounds across tissues, with the ultimate goal to identify functional and chemically similar compounds. The method was evaluated on 718 compounds in CMap that were profiled across three different cell lines. Compounds were compared to one another using the probabilistic score, with the external Anatomic Therapeutic Chemical (ATC) classification codes used to validate connectivity, where an ATC code of four indicates compounds share similar functionality (Parkkinen and Kaski, 2014).

4.1.3.4 cudaMap: a GPU accelerated program for gene expression connectivity mapping

The current and extended methods to query the CMap have failed to address the computational load for analysis, particularly with the repeated permutation tests. This issue is addressed by a study which harnessed the computational power of NVIDIA Graphics Processing Units (GPUs) to reduce the processing times for data analysis (McArt *et al.*, 2013) and presented their high-performance computing (HPC) models as a software platform called the cudaMAP. The authors further demonstrated that the ssCMap approach on this platform was capable of analysing multiple gene expression signatures in a matter of minutes rather than days. Although the cudaMAP is not a new method to query the CMap database, it addresses one of the significant issues surrounding computational drug repositioning, the computational demand.

4.1.3.5 SPIED: a searchable platform-independent expression database

The searchable platform-independent expression database (SPIED) compiles expression profiles from GEO and CMap to create an extensive database of expression correlations between diseases and

compounds (Williams, 2012). The database contains over 100,000 samples and includes experiments that lack controls. The author addresses the no control reference issue by generating fold change profiles using the experimental series average as a reference and uses a simple regression scoring scheme to match diseases to candidate compounds. The method essentially searches for anti-correlated disease signatures to candidate compounds in the CMap.

The authors initially validated their method by illustrating estrogen transcriptional response was anti-correlated to estrogen antagonists fulvestrant (R^2=-0.76), tamoxifen (R^2=-0.46) and raloxifene (R^2=-0.63). Further validation was performed by combining AD disease signatures from human and rodent brains, which identified 24 genes that were conserved in the disease. Using this small AD expression profile in the SPIED identified galantamine (R^2 = -0.34) as the 19th most significant result, a compound that is clinically prescribed to AD patients. However, the correlation of -0.34 can be interpreted as a weak-moderate negative correlation (Rumsey, 2011), and as the 19th most significant result from drug repositioning, may ordinarily be overlooked when querying a disease with unknown treatment. Furthermore, the published peer-reviewed portion of this thesis has identified hundreds of genes perturbed in AD (Patel, Dobson and Newhouse, 2019) and a simple 24 gene expression marker cannot account for all perturbations in AD and therefore would be inadequate for validation purposes.

4.1.4 Summary

Drug repositioning using the CMap database is potentially a useful approach to identify candidate compounds for disease intervention. However, many new methods focus on developing new, quicker and sophisticated techniques to query the CMap database, rather than addressing the quality of the initial data. The CMap has potential pitfalls in data quality, particularly the outdated processing methods used to QC the data, which needs to be addressed prior to developing algorithms to query the data. Furthermore, many methods lack robust computational validation, which can be evaluated by both drug-drug and disease-drug relations (Musa *et al.*, 2018). In drug-drug validation, a drug signature is queried through CMap to determine if the method retrieves compounds with similar chemical structure, while in disease-drug relations, a disease gene expression signature with treatment available in the CMap is queried to determine whether the method can identify candidate compounds that are already being clinically used to treat the disease in question.

This chapter attempts to address the current issues surrounding drug repositioning the CMap. First, the raw CMap data is re-processed using up-to-date robust pre-processing techniques. Next, inspired by previous research, different algorithms are developed to incorporate both up and down-regulated genes to search the reprocessed CMap database for candidate compound identification. The methods developed in this chapter will be validated using both drug-drug and disease-drug relations to identify

the best-equipped method for drug repositioning the CMap. Finally, two AD brain disease signatures identified in this thesis; one from AsymAD brains and the second from AD brains will be queried in the CMap to identify candidate compounds that may potentially intervene at both the asymptomatic and symptomatic stage of the disease. The work undertaken in this chapter is purely exploratory with all candidate compounds requiring further experimental investigations.

4.2 Method

4.2.1 The CMap data processing

The raw CMap intensity data (build 02) was manually downloaded from https://portals.broadinstitute.org/cmap/ in August 2016. The data consisted of small compounds expression profiled across five different cell lines and on three different Affymetrix expression chips (see background section "The Connectivity Map" for additional details).

Each cell line and expression chip was processed independently in RStudio (version 1.1.447) using R (version 3.4.4). The raw gene expression data were first pre-processed using the R package "gcrma" (version 2.50.0), which Robust Multi-Array (gcRMA) background corrects, log2 transforms, and then quantile normalises the data. Next, to account for technical and latent variation between batches due to the CMap experimental design consisting of small batches of 2-5 samples, batch effects were explored and adjusted by SVA using the R package "sva" (version 3.10.0). Drug-repositioning in this thesis was based on DE results, which is performed between two experimental groups. Therefore, compounds that lacked multiple expression profiles repeated at the same concentration were removed from further analysis.

Next, to ensure homogeneity within the control group, outlying samples were iteratively identified and removed using the fundamental network concepts described in (Oldham, Langfelder and Horvath, 2012). Then, to enable cross-platform probes to be comparable, Affymetrix platform-specific probe identifiers were annotated to their corresponding universal Entrez gene identifiers using the "hgu133a.db" R annotation file. Finally, DE analysis was performed using the R package "limma" (version 3.20.9) (Smyth, 2005).

4.2.2 Drug-Disease mapping

The Therapeutic Targets Database (TTD) is enriched with information on drug-disease relations, including compounds that are FDA approved to treat specific diseases (Yang *et al.*, 2016). The TTD was acquired in June 2017 and contained drug-disease mappings for 31,614 compounds. The compounds in the CMap database were annotated to the TTD to identify the diseases that the CMap compounds are currently FDA approved to treat.

4.2.3 Method development to query the CMap database

Four methods were developed to query the re-processed CMap database. Each method requires the disease gene expression signature in the format output from differential expression analysis. The general idea behind each method was to identify compounds that may "neutralise" the disease expression signature queried. Unlike the original CMap algorithm, the developed methods are not limited by 500 genes and allow all genes perturbed in the disease signature, including up and down-regulated genes, to be simultaneously queried in the CMap database to identify candidate compounds. A "candidate compound" is a compound deemed as a significant result by any method used to query the CMap. The four methods developed are "bi-directional enrichment", "anti-correlation", "ranked genes", and a "pathway-based" approach.

4.2.3.1 Method 1: Bi-directional enrichment method development

This method was inspired by the typical GSEA approach. It uses an enrichment test to assess the overlap between the DEG's in a disease signature and DEG's in every CMap compound, with the added modification of identifying inverse relation, i.e. matching compounds where the gene is significantly perturbed in the opposite direction to that from the disease.

First, the CMap reference database was prepared. For each compound, an "UP" was appended to the gene name if the gene's logFC was positive and a "DOWN" was appended to the gene name if the gene's logFC was negative. For instance, if gene "A" was up-regulated, and gene "B" was down-regulated than the gene names would be changed to "A_UP" and "B_DOWN" respectively. Then, the gene list for all CMap compounds was filtered to only those that were significantly differentially expressed (FDR adjusted p-value ≤ 0.05).

Next, the disease signature was prepared. The disease-gene list is first subset to only significant DEG's (FDR adjusted p-value ≤ 0.05) to create the disease signature profile. Next, similar to the CMap reference database, an "UP" and "DOWN" was appended to the gene names; however, as an anti-relation between disease and drug was required, the criteria was reversed. Thus, an "UP" was appended to a gene when the logFC was negative, and a "DOWN" was appended when the logFC was positive. For instance, if gene "A" was up-regulated, and gene "B" was down-regulated than the gene names would be changed to "A_DOWN" and "B_UP" respectively.

An enrichment test was performed on the two gene lists using the WGCNA (version 1.51) "userListEnrichment" function. This function requires three arguments ("geneR", "labelR" and "fnIn"). The "geneR" argument was a vector of all gene identifiers in the analyses, which includes both the DEG's and background genes that are not perturbed by the drug. Essentially, this was a list of all genes in the CMap database. The "labelR" argument was a list of all DEG's within a single compound, and the

"fIIn" argument was the DEG's in the disease signature. The enrichment test was repeated for each compound in the CMap using the R packages "foreach" (version 1.4.4) and "doParallel" (version 1.0.11) for parallel processing, and results were corrected for multiple testing using FDR.

4.2.3.2 Method 2: Anti-correlation method development

This method was inspired by SPIED. It used a simple anti-correlation approach to identify compounds where the expression signature is anti-correlated with the disease expression signature. First, the CMap reference database was created. The logFC for all genes were converted to +1 if the logFC value was positive, or -1 if the logFC value was negative. Next, the disease query signature was prepared. The disease genes are first subset to those that are differentially expressed (FDR adjusted p-value ≤ 0.05). Then, the logFC values are converted to +1 or -1 using a reversed criteria to that used in the CMap database, i.e. the logFC was converted to +1 if the logFC value was negative, or -1 if the logFC value was positive.

Then, using the R "cor.test" function, a Spearman's correlation test was then iteratively performed between the disease signature and every compound expression signature in the CMap database. This method was repeated for each compound in the CMap database and results were corrected for multiple testing using FDR.

4.2.3.3 Method 3: Ranked gene-based method development

This method was inspired by the ssCMap algorithm. It incorporates the extent of statistical perturbation (p-value) in combination with the magnitude of differential expression (logFC) of the disease gene expression signature when identifying candidate compounds. First, the reference CMap database was created. The genes for each CMap compound were ordered from the most significantly up-regulated gene to the most significantly down-regulated gene. Next, the disease gene expression signature was created. The significant DEG's (FDR adjusted p-value ≤ 0.05) from the disease gene expression signature was ordered using a reverse criteria to that used in the CMap reference database so that the most significantly down-regulated gene is at the top of the list and the most significantly up-regulated gene is at the bottom of the list.

The Rank-Rank Hypergeometric Overlap (RRHO) test was then used through the R package "RHHO" (version 1.18.0) to assess the significant overlap between the disease gene expression signature and the CMap reference database. The RRHO algorithm steps through the two gene lists and measures the statistical significance of the number of overlapping genes through a hypergeometric enrichment of *k* overlapping genes calculated for all possible threshold pairs (Plaisier *et al.*, 2010). The "RRHO" function incorporates a 1000 permutation tests when calculating the p-value. This method was repeated for

each compound using the R packages "foreach" (version 1.4.4) and "doParallel" (version 1.0.11) for parallel processing the CMap database.

4.2.3.4 Method 4: Pathway-based method development

This method uses a system-biology approach to first asses the biological pathways perturbed in the disease gene expression signature, then attempts to identify candidate compounds that may intervene with the same biological pathways with a reversed effect. For instance, if a disease gene expression signature is found to "activate" the "glutamate-glutamine cycle", then a compound that inhibits the same biological pathway is sought after. Therefore, this method is not limited by individual gene overlap and looks at the bigger picture of how interactions between different genes may influence biological processes and uses this information to create disease-drug relations.

The method uses the signalling pathway impact analysis (SPIA) to identify biological pathways perturbed by each compound. The SPIA method was implemented through the R package "SPIA" (version 2.30.0) and uses the human signalling pathways from KEGG (Ogata *et al.*, 1999), which contains pathway activation and repression information that was derived from gene and protein signalling and interactions (Tarca *et al.*, 2009). The SPIA method combines the evidence obtained from the classical enrichment analysis with a novel topology type of evidence to infer actual perturbation of a given pathway (Tarca *et al.*, 2009).

The pathway-based drug repositioning approach begins by creating a reference CMap database. First, the SPIA method is iteratively performed on the DEG's of every compound in the CMap database to generate a list of significant ($p \leq 0.05$) biological pathways "activated" and "inhibited" by each compound. Next, the disease gene expression signature was analysed with SPIA to identify significant ($p \leq 0.05$) biological pathways "activated" and "inhibited" in the disease. The biological pathway perturbations are reversed in the disease signature so "activated" pathways are changed to "inhibited" and vice-versa. A standard hypergeometric test using the R function "phyper" was performed on pathways labels, which essentially assess the overlap of the number of same pathways being perturbed in the disease and by the compound. The resulting p-value is corrected for multiple testing using FDR. The output of this method specifies the number of biological pathways perturbed in the disease and a list of compounds with respective biological pathways it may correct with a corresponding p-value.

4.2.4 CMap database query algorithm validations

The four methods developed to query the CMap database were validated through two recommended techniques; drug-drug relations and disease-drug relations (Musa *et al.*, 2018). The drug-drug relation queries a drug signature to identify compounds of similar properties, while the disease-drug relation

queries a known disease gene expression signature to identify compounds known to interact with the disease in question. The drug-rug approach was adapted for this thesis and both are described below.

4.2.4.1 Drug-drug relation

Five compounds were randomly selected from the CMap database using the "sample" function in R. The five compounds had their gene expression signatures reversed by inverting the logFC signs for all genes. The compounds were thereafter treated as "dummy" disease gene expression signatures and queried in the CMap database using the four developed methods ("bi-directional enrichment", "anti-correlation", "ranked genes", and "pathway-based). The underlying hypothesis behind this strategy expects the "dummy" disease expression signatures, which have originated from compounds in the CMap, should be a perfect match to the same compound in the CMap database, and each method should identify this compound as the most statistically significant result, followed by compounds with similar properties. This process can be used to validate that the methods are capable of identifying compounds where the gene expression signature is a "reverse" match to queried disease gene expression signature.

4.2.4.2 Disease-drug relation

Disease datasets where the FDA approved treatment was available in the CMap database were sought from public repositories ArrayExpress and GEO. The criteria used to identify potential datasets were: 1) data must be generated from the Affymetrix or Illumina microarray expression platform, 2) data must be generated from a human cell line, 3) dataset must include both diseased and complimentary healthy controls in the same experimental match, allowing for differential expression analysis to be performed, 4) dataset must contain at least 10 cases and 10 controls and 5) dataset description or associated publication must explicitly mention the patients were not on any disease-modifying treatment prior to sample extraction.

Microarray gene expression studies were acquired from public repositories using the R package "ArrayExpress" (version 1.38.0). Raw gene expression data generated on the Affymetrix platform were "gcRMA" background corrected using R package "gcRMA" (version 2.50.0), log2 transformed and then quantile normalised. Datasets generated on the Illumina platform were "normexp" background corrected, log2 transformed, and quantile normalised using the "limma" R package (version 3.20.9).

Sex was then predicted using the R package "massiR" (version 1.0.1) and subjects with discrepancies between predicted and recorded sex removed from further analysis. Next, within each gender and disease diagnosis group, probes above the "X" percentile of the log2 expression scale in over 80% of the samples were deemed "reliably detected". To account for the variation of redundant probes across different BeadArrays, the "X" percentile threshold value was manually adjusted until a variety of robust

literature defined house-keeping genes were correctly defined as expressed or unexpressed in their corresponding gender groups (Chang *et al.*, 2011). Any probe labelled as "reliably detected" in any group (based on gender and diagnosis) was taken forward for further analysis from all samples within that dataset. This essentially eliminates noise (Lazar *et al.*, 2013) and ensures disease and gender-specific signals are captured within each dataset.

Publicly available data is often accompanied by a lack of sample processing information, making it impossible to adjust for systematic errors introduced when samples are processed in multiple batches, a term often known as "batch effects". To account for both known and possible latent variation, batch effects were estimated and removed by SVA using R package "sva" (version 3.10.0). Next, to ensure homogeneity within biological groups, outlying samples were iteratively identified and removed using the fundamental network concepts described in (Oldham, Langfelder and Horvath, 2012), followed by annotating platform-specific probe ID's to their corresponding universal Entrez gene identifiers using the appropriate R annotation files; "hgu133plus2.db", "hugene10sttranscriptcluster.db" and "illuminaHumanv4.db". Finally, differential expression analysis was performed using the R package "limma" (version 3.20.9) to create the expression signature for each disease, which was queried in the CMap database using the four developed methods ("bi-directional enrichment", "anti-correlation", "ranked genes", and "pathway-based").

The underlying hypothesis behind this validation strategy assumes querying known disease gene expression signatures should identify the CMap compounds that are clinically used to treat the disease.

4.2.5 Identifying candidate compounds for AD

This thesis explored transcriptomic changes in probable AsymAD brains and identified significant brain changes in the FC of AsymAD subjects when compared to healthy controls, suggesting this may be the first brain region affected in AD. Further analysis identified a gene expression signature consisting of 398 genes, which was deemed to represent early changes in AD, prior to clinical symptoms. In addition, this thesis identified a gene expression signature in the TC of AD subjects through the largest AD meta-analysis known to date. This gene expression signature consisted of 323 genes and represented robust changes in AD brains.

The AsymAD and AD disease gene expression signatures were independently queried through the re-processed CMap database to identify candidate compounds that may intervene with AD in the asymptomatic and symptomatic phase of the disease. The compounds identified are merely candidate compounds and would need further experimental validation.

4.2.6 Predicting BBB permeability

Compounds of interest were queried in the web-based database "admetSAR2.0" which is available at http://lmmd.ecust.edu.cn/admetsar2/. The application uses 210,000 experimental data for 96000 compounds and 27 computational models to predict chemical and biological properties of compounds, including BBB permeability (Yang *et al.*, 2019). Compounds of interest identified by the drug repositioning pipelines were manually analysed in the ametSAR2.0 application to assess the compounds BBB permeability likelihood.

4.2.7 Data availability

All data processing scripts are available at https://doi.org/10.5281/zenodo.3518008

4.3 Results

4.3.1 Summary of processing the CMap data

The CMap data consisted of 7056 raw intensity files, of which 6100 were generated from compounds and 956 generated from "control vehicles". A summary of sample numbers by cell line and expression chip is provided in Table 4.1.

Table 4.1: Summary of CMap sample numbers by cell line and expression chip

Expression GeneChip	HG U133A					Total
Cell Line	HL60	MCF7	PC3	SKMEL5	ssMCF7	
Compounds	344	171	124	17	18	**674**
Control vehicle	52	47	24	5	5	**133**

Expression GeneChip	HT-HG U133A					
Cell Line	HL60	MCF7	PC3	SKMEL5	ssMCF7	
Compounds	885	2740	1617	0	0	**5242**
Control vehicle	125	409	253	0	0	**787**

Expression GeneChip	HT-HG U133A EA					
Cell Line	HL60	MCF7	PC3	SKMEL5	ssMCF7	
Compounds	0	184	0	0	0	**184**
Control vehicle	0	36	0	0	0	**36**

Total	1406	3587	2018	22	23	7056

The CMap data was generated across three expression chips (HG-U133A, HT-HG U133A, HT-HG U133A EA), five cell lines (HL60, MCF7, PC3, SKMEL5, ssMCF7) and consisted of expression profiles from 6100 compounds and 956 control vehicles. The MCF7 cell line that was expression profiled on the HT-HG U133A expression chip contained the largest number of samples. The cell lines are MCF7 (human breast epithelial adenocarcinoma cell line), ssMCF7 (MCF7 cell lines cultured in phenol red-free DMEM), HL60 (human promyelocytic cell line), PC3 (epithelial cell line from human prostate adenocarcinoma) and SKMEL5 (human malignant melanoma cell line).

DE analysis calculates the average expression difference between two experimental groups, where each group must contain at least two samples. Therefore, prior to DE analysis, a summary of compounds that were repeated in each cell line and expression chip was compiled and is presented in Table 4.2. Compounds with multiple expression profiles were discovered to have been experimentally repeated 2-16, 17, 19, 22, 36, or 64 times. Multiple samples were also discovered to have been experimentally repeated at different concentrations on the same cell line and expression chip, each generating different expression profiles for the same compound. These expression profiles were treated as unique instances, and the compound names were modified to include concentration details to aid in differentiation. For instance, compound trichostatin was expression profiled at two different concentrations; 1e-06 and 1e-07, therefore the names of the two compounds have been converted to trichostatin_1e-06 and trichostatin_1e-07 respectively.

The MCF7 cell line from the HT HG U133A expression chip contained the highest number of unique compounds and control vehicles that were experimentally repeated multiple times. Therefore, only this cell line and expression chip were used for further analysis.

Table 4.2: Summary of the number of repeated compounds by cell line and expression chip

Expression GeneChip	Cell line	N.o. Repeats	
		0	>=1
HG U133A	HL60	351	3
	MCF7	78	40
	PC3	109	7
	SKMEL5	17	0
	ssMCF7	14	2
HT HG U133A	HL60	741	34
	MCF7	95	1146
	PC3	853	337
	SKMEL5	0	0
	ssMCF7	0	0
HT HG U133A EA	HL60	0	0
	MCF7	42	54
	PC3	0	0
	SKMEL5	0	0
	ssMCF7	0	0

This table summarises the number of compounds that were experimentally repeated in each cell line and expression chip. The cell lines are MCF7 (human breast epithelial adenocarcinoma cell line), ssMCF7 (MCF7 cell lines cultured in phenol red-free DMEM), HL60 (human promyelocytic cell line), PC3 (epithelial cell line from human prostate adenocarcinoma) and SKMEL5 (human malignant melanoma cell line). The CMap compounds may have been experimentally repeated 2-16, 17, 19, 22, 36, or 64 times. The MCF7 cell line from the HT HG U133A expression chip has the greatest number of unique compounds and control vehicles. Therefore, only this cell line and expression chip was used for further analysis

4.3.2 Summary of differential expression analysis

Following QC, the MCF7 cell line from the HT-HG U133A expression chip contained 12,501 annotated probes and 1146 unique compound expression profiles. Differential expression analysis was performed using each compound repeated at the same concentration against the pooled control vehicles. Differential expression analysis on the CMap compounds identified a range of 26-9167, a mean of 1157.5 and an interquartile range (IQR) of 312-1576.5 (Q1-Q3) DEG's across the 1146 compounds.

4.3.3 Drug-disease mapping identifies 768 unique drug-disease relations

The TTD contained information on 31,614 drug-disease relations and were used to identify diseases which can be treated by any of the 1146 CMap compounds. The CMap database contains 1121 unique compounds, with some compounds experimentally repeated at different concentrations and therefore, accounts for 1146 expression profiles. Only 612 CMap compounds annotated to 299 different diseases; however, due to multiple drug-disease combinations, 768 CMap drug-disease relations were identified and are provided in Supplementary Table 4.1. The top ten diseases targeted by multiple CMap compounds are illustrated in Figure 4.1

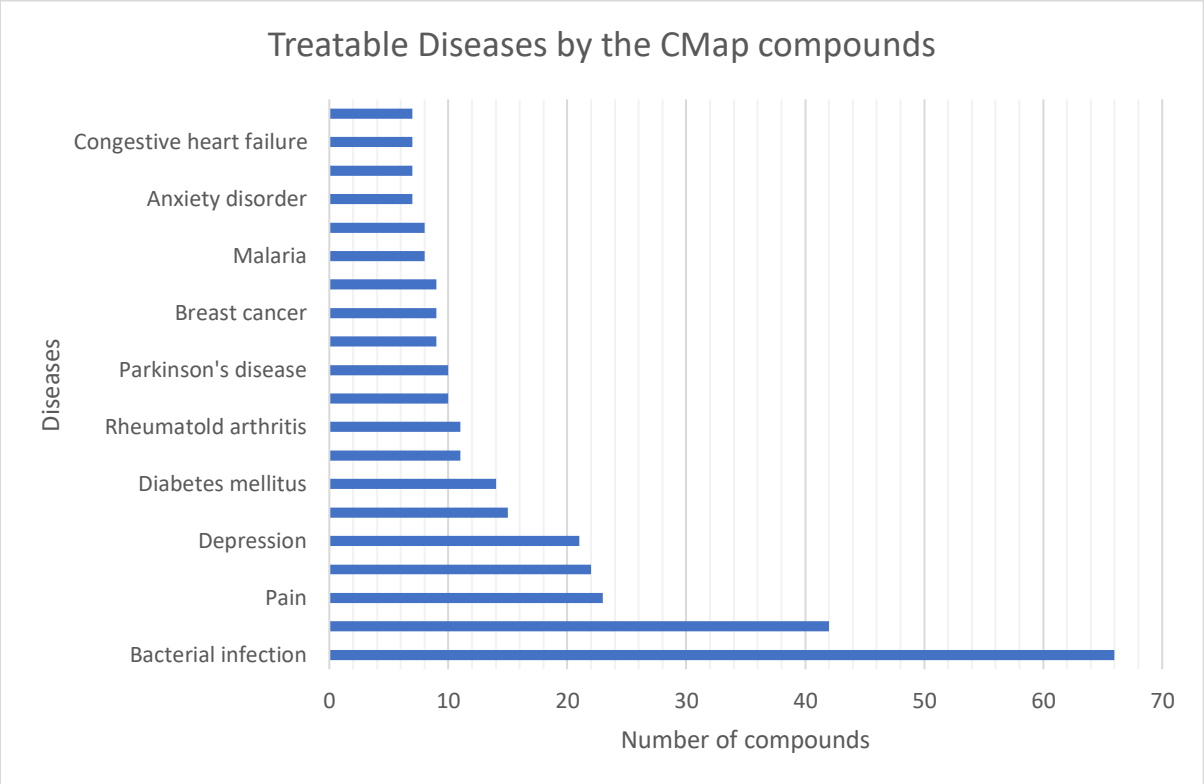


Figure 4.1: Top 10 most treatable diseases by multiple CMap compounds. The compounds in the CMap database were annotated to the disease for which, according to TTD, are FDA approved to treat. A total of 768 Drug-Disease relations were identified in the CMap database. The CMap database contains the most number of compounds to treat bacterial infection.

4.3.4 Drug-drug validation results

The Five random compounds identified from the CMap database were iopanoic acid, atractyloside, ursolic acid, methoxamine and 6-azathymine. These compounds had their DE logFC signals reversed, were then treated as “dummy” disease signatures and queried in the CMap database using the four different methods developed in this thesis. The five most significant results from each method are provided in Table 4.3. All four methods successfully identify the same queried compound as the most

Table 4.3: Top five significant results from Drug-Drug validation analysis

Compound (DEG's)	Method 1		Method 2		Method 3		Method 4	
	CMap compound	p-value	CMap compound	p-value	CMap compound	p-value	CMap compound	p-value
Iopanoic acid (263)	iopanoic acid_7e-06	0.00E+00	iopanoic acid_7e-06	0.00E+00	iopanoic acid_7e-06	0.00E+00	iopanoic acid_7e-06	4.11E-09
	procarbazine_1.56e-05	4.35E-52	10-methoxyharmalan_1.86e-05	7.21E-04	10-methoxyharmalan_1.86e-05	0.00E+00	noscapine_9.6e-06	1.33E-04
	piribedil_1.2e-05	1.62E-50	scopolamine N-oxide_1e-05	3.24E-03	acemetacin_9.6e-06	0.00E+00	lisinopril_9e-06	2.75E-03
	S-propranolol_1.36e-05	2.13E-49	naproxen_1.74e-05	4.08E-03	alfaxalone_1.2e-05	0.00E+00	albendazole_1.5e-05	3.09E-03
	tribenoside_8.4e-06	1.11E-40	talampicillin_7.8e-06	1.07E-02	amitriptyline_1.28e-05	0.00E+00	bufexamac_1.8e-05	3.45E-03
Atractyloside (293)	atractyloside_5e-06	0.00E+00	atractyloside_5e-06	0.00E+00	atractyloside_5e-06	0.00E+00	atractyloside_5e-06	5.42E-11
	guanadrel_7.6e-06	7.32E-49	calcium folinate_7.8e-06	1.64E-03	AG-013608_1e-05	0.00E+00	cefazolin_8.4e-06	3.41E-05
	merbromin_5e-06	9.50E-49	ethoxyquin_1.84e-05	5.81E-03	alpha-ergocryptine_7e-06	0.00E+00	vitexin_9.2e-06	6.05E-05
	sulfadimethoxine_1.28e-05	5.88E-40	alpha-ergocryptine_7e-06	7.35E-03	atractyloside_5e-06	0.00E+00	remoxipride_9.8e-06	9.94E-05
	glycopyrronium bromide_1e-05	1.81E-37	bendroflumethiazide_9.4e-06	1.13E-02	calcium folinate_7.8e-06	0.00E+00	flufenamic acid_1.42e-05	4.46E-04
Ursolic acid (792)	ursolic acid_8.8e-06	0.00E+00	ursolic acid_8.8e-06	0.00E+00	ursolic acid_8.8e-06	0.00E+00	ursolic acid_8.8e-06	4.77E-37
	diazoxide_1.74e-05	6.91E-177	iopamidol_5.2e-06	6.07E-21	0179445-0000_1e-05	0.00E+00	mephentermine_9.4e-06	7.52E-12
	cefotaxime_8.4e-06	8.39E-173	tetrandrine_6.4e-06	1.94E-10	0179445-0000_1e-06	0.00E+00	prochlorperazine_6.6e-06	4.34E-09
	clindamycin_8.6e-06	2.61E-167	podophyllotoxin_9.6e-06	3.76E-09	0297417-0002B_1e-05	0.00E+00	hydrocortisone_1.1e-05	7.90E-09
	doxycycline_8.4e-06	1.19E-155	dihydroergocristine_5.6e-06	3.28E-08	10-methoxyharmalan_1.86e-05	0.00E+00	perphenazine_1e-05	4.04E-08

Methoxamine (222)	methoxamine_1.62e-05	0.00E+00	methoxamine_1.62e-05	0.00E+00	methoxamine_1.62e-05	0.00E+00	methoxamine_1.62e-05	3.60E-03
	cefaclor_1.04e-05	3.17E-43	phenacetin_2.24e-05	4.46E-08	6-azathymine_3.14e-05	0.00E+00	amitriptyline_1.28e-05	1.08E-02
	pyrazinamide_3.24e-05	4.09E-40	L-methionine sulfoximine_2.22e-05	1.39E-04	acebutolol_1.08e-05	0.00E+00	oxamniquine_1.44e-05	1.08E-02
	lovastatin_9.8e-06	3.66E-36	liothyronine_6.2e-06	1.78E-04	apramycin_7.4e-06	0.00E+00	atropine oxide_1.18e-05	1.44E-02
	betahistine_1.72e-05	6.82E-29	cefaclor_1.04e-05	4.51E-04	arcaine_1.48e-05	0.00E+00	metoprolol_5.8e-06	1.44E-02
6-azathymine (324)	6-azathymine_3.14e-05	0.00E+00	6-azathymine_3.14e-05	0.00E+00	6-azathymine_3.14e-05	0.00E+00	6-azathymine_3.14e-05	1.55E-18
	phenacetin_2.24e-05	3.69E-20	amiprilose_1.18e-05	8.84E-06	0179445-0000_1e-06	0.00E+00	flumequine_1.54e-05	1.54E-06
	molsidomine_1.66e-05	1.73E-16	phenacetin_2.24e-05	2.10E-05	0317956-0000_1e-05	0.00E+00	(+)-isoprenaline_1.1e-05	1.33E-04
	omeprazole_1.16e-05	3.68E-15	proglumide_1.2e-05	2.78E-05	0317956-0000_1e-06	0.00E+00	ethaverine_9.2e-06	7.22E-04
	cinchocaine_1.16e-05	1.71E-14	dicoumarol_1.18e-05	2.60E-04	6-azathymine_3.14e-05	0.00E+00	idoxuridine_1.12e-05	1.10E-03

Five random compounds were identified from the CMap database; iopanoic acid, atractyloside, ursolic acid, methoxamine and 6-azathymine. These compounds had their differential logFC signals reversed and were then queried in the CMap database using four different methods. Some CMap compounds were experimentally used multiple times at different concentration, generating different expression profiles for the same compound. Therefore, each compound has the concentration at which it was used appended to the end of the compound name. Method 1 represents the Bi-directional enrichment approach. Method 2 represents the anti-correlation approach. Method 3 represents the ranked genes approach, and method 4 represents the pathway-based approach. All four methods successfully identify the same queried compound as the most significant result, suggesting the methods are all capable of identifying compounds where the gene expression signature is a "reverse" match to queried disease gene expression signature.

significant result, suggesting all four methods can identify compounds where the gene expression signature is a "reverse" match to queried disease gene expression signature. However, as shown in Table 4.4, the processing time and the number of additional candidate compounds identified by each method varies, with method 3 taking considerably much longer than the remaining methods. In addition, a positive correlation between the number of DEG's in the query signature and the number of candidate compounds identified, and the execution time for the method is observed. On average, the bi-directional enrichment method identifies the greatest number of candidate compounds, followed by the ranked genes, anti-correlation and then the pathway-based method.

Table 4.4: Summary of significant hits from drug-drug validation analysis

Drug	N.o. DEG	Method 1		Method 2		Method 3		Method 4	
		N.o Candidate compounds	Execution time (s)	N.o Candidate compounds	Execution time (s)	N.o Candidate compounds	Execution time (s)	N.o Candidate compounds	Execution time (s)
iopanoic acid	263	863	35.71	12	0.62	223	1055.94	39	198.01
atractyloside	293	891	44.76	18	0.70	177	1195.95	122	207.49
ursolic acid	792	1035	42.80	392	0.91	833	4185.68	252	213.57
methoxamine	222	750	37.77	36	0.55	232	945.00	44	205.80
6-azathymine	324	830	41.44	81	0.51	198	1432.68	40	211.87
Average	379	874	40.50	108	0.66	333	1763.05	99	207.35
Correlation Coefficient		0.90	0.44	0.99	0.88	0.97	1.00	0.93	0.64

Five random compounds were identified from the CMap database; iopanoic acid, atractyloside, ursolic acid, methoxamine and 6-azathymine. These compounds had their differential expression logFC signals reversed and were then queried in the CMap database using the four different methods. Method 1 represents the Bi-directional enrichment approach; method 2 represents the anti-correlation approach, method 3 represents the ranked genes approach and method 4 represents the pathway-based approach. The number of significant results and the execution time from each method is provided above. A positive correlation of the number of DEG's in the query signature and the number of significant compounds is observed.

4.3.5 Disease-drug validation results

Three disease datasets were identified in the public domain where corresponding FDA approved treatment was available in the CMap database. The diseases were tuberculosis, breast cancer and prostate cancer. Each dataset was independently processed, and an overview of the dataset demographics after processing is provided in Table 4.5. Differential expression analysis was independently performed in each dataset to identify DEG's, which represented the diseases gene expression signature that was queried in the CMap database. The five most significant candidate compounds identified from this analysis is provided in Table 4.6, and a summary of all candidate compounds is provided in Table 4.7.

Table 4.5: Demographics of the datasets used in the disease-drug validation analysis

	Tuberculosis	Breast cancer	Prostate cancer
Accession number	E-GEOD-34151	E-GEOD-54002	E-GEOD-29079
Platform	Illumina	Affymetrix	Affymetrix
Expression chip	HT-12 v4.0	U133 plus 2.0	Exon 1.0 ST
Tissue source	Primary dendritic cells	Mammary gland	Prostate epithelium
N.o. cases	124	405	46
N.o. controls	112	15	47
N.o. DEG's identified	2947	5628	6674

Three disease datasets (tuberculosis, prostate cancer and breast cancer) were identified from the public repository ArrayExpress where treatment was available in the CMap database. Each dataset was independently processed to identify DEG's, which represent the diseases gene expression signature. Sample numbers are provided after QC.

The bi-directional enrichment approach (method1) correctly identified FDA approved drugs to treat the cancer gene expression signatures, with two cancer drugs being the two most significant result when querying the breast cancer disease gene expression signature. However, this method failed to identify any FDA approved drugs for tuberculosis and prostate cancer.

The anti-correlation approach (method 2) was the only method to consistently identify FDA approved drugs to treat all three diseases queried, although they were never ranked in the top ten most significant results. This method also identified the highest number of candidate compounds when compared to the remaining three methods, particularly when querying the prostate and breast cancer disease gene expression signatures where 886 and 762 candidate compounds were identified respectively. This is more than 66% of the total number of compounds available in the CMap database (total = 1146 compounds), and therefore statistically, it is highly likely an FDA approved compound used to treat the disease would be identified by chance.

In contrast, the rank gene-based approach was more reserved and identified 1, 3 and 4 candidate compounds when querying tuberculosis, prostate cancer and breast cancer disease gene expression signatures respectively. However, none of the candidate compounds is FDA approved to treat any of the diseases, although the single drug identified for tuberculosis, a bacterial infection, is an anti-bacterial compound. However, no relation between the candidate compound and tuberculosis is mentioned in the TTD, and it cannot be ignored anti-bacterial compounds are the most abundant in the CMap database. Nevertheless, if this anti-bacterial compound can be used to treat tuberculosis, method 3 would be the most sensitive method.

Table 4.6: Top five significant results from the disease-drug validation analysis

Disease	Method 1			Method 2			Method 3			Method 4		
	CMap compound	**FDR p-value**	**Current prescription**	**CMap compound**	**FDR p-value**	**Current prescription**	**CMap compound**	**FDR p-value**	**Current prescription**	**CMap compound**	**FDR p-value**	**Current prescription**
Tuberculosis	cephaeline	9.36E-40	<NA>	GW-8510	6.75E-10	<NA>	cefsulodin	9.00E-03	Bacterial infection	nortriptyline	2.42E-04	Depression
	emetine	1.48E-27	Hepatitis virus infection	H-7	2.42E-09	<NA>				prasterone	2.42E-04	Chronic obstructive pulmonary disease
	sirolimus	2.19E-17	Dutch elm disease; Organ rejection	mitoxantrone	4.84E-09	Cancers				liothyronine	4.65E-04	Hypothyroidism
	tanespimycin	7.06E-17	Breast cancer; Multiple myeloma	doxorubicin	1.80E-07	Cancers				tracazolate	4.97E-04	<NA>
	puromycin	1.99E-16	<NA>	0179445-0000	1.89E-07	<NA>				hycanthone	1.05E-03	<NA>
Prostate cancer	LY-294002	5.05E-42	<NA>	chlorpropamide	7.17E-75	Non-insulin dependent diabetes; Diabetes mellitus	ritodrine	0.00E+00	Premature labour	lisuride	7.79E-05	Parkinson's disease
	cephaeline	9.15E-41	<NA>	disopyramide	8.79E-52	Ventricular arrhythmias; Urinary incontinence	disopyramide	4.00E-02	Urinary incontinence; Ventricular arrhythmias	meticrane	3.22E-04	<NA>
	emetine	1.31E-39	Hepatitis virus infection	clobetasol	5.88E-50	Inflammation and itching	thiamphenicol	4.00E-02	Bacterial infection	chlortalidone	7.74E-04	Hypertension
	tretinoin	3.17E-23	Acne vulgaris; Acute promyelocytic leukaemia	sulfamethoxazole	5.78E-45	Infectious disease				labetalol	2.15E-03	Hypertension
	nocodazole	2.45E-22	<NA>	procyclidine	1.64E-43	Parkinson's disease				amikacin	3.44E-03	Bacterial infection

Breast cancer	geldanamycin	1.75E-50	Kidney cancer; Peripheral nerve damage	Prestwick-1080	4.69E-87	<NA>	pheneticillin	0.00E+00	Bacterial infection	lisuride	7.27E-04	Parkinson's disease
	tanespimycin	8.76E-50	Breast cancer; Melanoma; Multiple myeloma	phenelzine	9.44E-50	Depression; Anxiety disorder	trichostatin A	0.00E+00	anti-fungal	isotretinoin	7.62E-04	Acne vulgaris
	withaferin A	4.75E-40	<NA>	rifabutin	1.13E-48	Tuberculosis	flucloxacillin	5.00E-03	Bacterial infection	cefuroxime	9.31E-04	Bacterial infection
	cephaeline	8.42E-39	<NA>	meptazinol	1.11E-41	Pain	isometheptene	4.30E-02	headaches	amikacin	1.72E-03	Bacterial infection
	15-delta prostaglandin J2	5.66E-36	<NA>	Prestwick-984	3.60E-38	<NA>				naftidrofuryl	3.29E-03	Vascular disorder

Three disease datasets (tuberculosis, prostate cancer and breast cancer) were identified from the public repository ArrayExpress where treatment was available in the CMap database and where the human samples were not on any disease-modifying treatment. Each dataset was independently processed and analysed to identify the DEG's, which represented the diseases gene expression signature. These disease gene expression signatures were queried in the CMap database using the four different methods developed. "N.o. prescribed drugs identified" represents candidate compounds identified by the method that has been FDA approved to treat the queried disease according to the TTD. "<NA>" represents a drug that is not FDA approved to treat diseases according to TTD. Method 1 represents the Bi-directional enrichment approach; method 2 represents the anti-correlation approach, method 3 represents the ranked genes approach and method 4 represents the pathway-based approach. Compounds highlighted in red indicate compounds that are FDA approved to treat the disease queried.

Table 4.7: Summary of disease-drug validation

Disease	Method 1		Method 2		Method 3		Method 4	
	N.o. compounds identified as significant	N.o. prescribed drugs identified	N.o. compounds identified as significant	N.o. of prescribed drugs identified	N.o. compounds identified as significant	N.o. of prescribed drugs identified	N.o. compounds identified as significant	N.o. prescribed drugs identified
Tuberculosis	66	0	171	6	1	0	106	2
Prostate cancer	73	1	886	4	3	0	99	0
Breast cancer	74	5	762	3	4	0	67	0
Average	71.00	2.00	606.33	4.33	2.67	0.00	90.67	0.67

Three disease datasets (tuberculosis, prostate cancer and breast cancer) were identified from the public repository ArrayExpress where treatment was available in the CMap database and where the human samples were not on any disease-modifying treatment. Each dataset was independently processed and analysed to identify the DEG's, which represented the diseases gene expression signature. These disease gene expression signatures were queried in the CMap database using the four different methods developed. "N.o. prescribed drugs identified" represents candidate compounds identified by the method that has been FDA approved to treat the queried disease according to the TTD. Method 1 represents the Bi-directional enrichment approach, method 2 represents the anti-correlation approach, method 3 represents the ranked genes approach, and method 4 represents the pathway-based approach

The final method to identify candidate compounds was a pathway-based approach (method 4), which correctly identified FDA approved drugs for tuberculosis but failed to identify FDA approved drugs for the cancer datasets correctly. The current results from this disease-drug analysis are inconclusive to suggest which method is best suited for identifying candidate compounds from a disease gene expression signature. However, method 2 can be regarded as the worst performing method based on the high number of candidate compounds identified in each analysis, indicating increased false positives. Therefore, this method was dropped from further use.

4.3.6 Candidate compounds identified for AsymAD treatment

The AsymAD gene expression disease signature was queried in the CMap database. The five most significant results from each method are provided in Table 4.8. Querying the AsymAD disease signature identified 3, 2 and 15 candidate compounds using methods 1, 3, and 4 respectively. No overlap of candidate compounds was identified across the four methods. The bi-directional enrichment approach (method 1) identified three candidate compounds with anisomycin, an antibiotic, being the most significant candidate compound. The ranked-gene based approach identified only two candidate compounds; tacrine (FDR adjusted p-value=0.03), an FDA approved drug to treat AD, and epirizole (FDR adjusted p-value=0.04), an anti-inflammatory drug. In addition, galantamine, another FDA approved compound to treat AD, identified as the 6th most significant compound, although it did not pass multiple corrections (FDR adjusted p-value=0.15).

The pathway-based approach identified 15 candidate compounds based on perturbation of biological pathways. Similar to the bi-enrichment based approach, the most significant candidate compound was an antibiotic (bacampicillin, FDR adjusted p-value=0.001). This method also identified thioperamide as the 5th most significant candidate compound (FDR adjusted p-value=0.02), which is an FDA approved drug to treat cognitive impairment.

4.3.7 Candidate compounds identified for AD treatment

Similar to the AsymAD, the AD gene expression disease signature was queried in the CMap database to identify candidate compounds. The five most significant results are also provided in Table 4.8. Only method 3 (ranked gene-based) and method 4 (pathway-based) identified candidate compounds for AD. The ranked gene-based approach identified 63 candidate compounds, with paromomycin, an antibiotic, being the most significant result (FDR adjusted p-value=0). This method also identified memantine; an FDA approved compound to treat AD, as the 62nd most significant compound (FDR adjusted p-value=0.04).

The pathway-based approach identified the activation of "bacterial invasion of epithelial cells" pathway in AD. This biological pathway was the only pathway identified as significantly perturbed in AD. Twelve candidate compounds were identified to inhibit this pathway with kawain; an FDA approved compound to treat cardiovascular disease, identified as the most significant (FDR adjusted p-value=0.02) candidate compound. However, according to the results, this compound significantly perturbs an additional three biological pathways (salmonella infection, focal adhesion, neuroactive ligand-receptor interaction) that are not affected in AD. The 5th most significant candidate compound identified by this pathway-based approach was velnacrine, an FDA approved compound to treat cognitive impairment.

Table 4.8: Top five most significant candidate compounds identified for Alzheimer's disease

Disease	Method 1			Method 3			Method 4		
	CMap compound	FDR p-value	Current prescription	CMap compound	p-value	Current prescription	CMap compound	FDR p-value	Current prescription
AsymAD	anisomycin	3.32E-02	Antibiotic	tacrine	3.10E-02	Alzheimer's disease	bacampicillin	1.54E-03	Bacterial infection
	SR-95531 (Gabazine)	3.32E-02	<NA>	epirizole	3.50E-02	Anti-inflammatory	cinchocaine	6.96E-03	<NA>
	cephaeline	3.57E-02	<NA>			<NA>	estropipate	6.96E-03	Hypogonadism
							metamizole sodium	8.88E-03	<NA>
				*galantamine	1.48E-01	Alzheimer's disease	thioperamide	1.80E-02	Cognitive impairment
AD				paromomycin	0.00E+00	Acute and chronic intestinal amebiasis	kawain	1.44E-02	Cardiovascular disease
				bendroflumethiazide	1.00E-03	High blood pressure	budesonide	2.16E-02	Asthma; Non-infectious rhinitis
				pyrantel	1.00E-03	Worm infections	neostigmine bromide	2.16E-02	<NA>
				ethoxyquin	2.00E-03	<NA>	phenelzine	2.16E-02	Depression; Anxiety disorder
				oxprenolol	4.00E-03	Angina; Hypertension	clomipramine	2.52E-02	Depression
				*memantine	4.40E-02	Alzheimer's disease	**velnacrine	3.00E-02	Cognitive impairment

*This thesis explored transcriptomic changes in probable AsymAD brains and identified significant brain changes in the FC of AsymAD subjects when compared to healthy controls, suggesting this may be the first brain region affected in AD. The 398 gene expression signature identified in the FC of AsymAD subjects represented early changes in the disease and was queried through the reprocessed CMap database to identify candidate compounds that may intervene with AD in the early stage of the disease, prior to clinical symptoms. In addition, this thesis identified a 323 gene expression signature in the TC of AD subjects through the largest AD meta-analysis known to date. This gene expression signature represented robust changes in AD brains and was queried through the reprocessed CMap database to identify candidate compounds that may intervene with AD once clinical signs of AD are present. The top five most significant candidate compounds identified by each query method is provided in the above table. Method 1 represents the Bi-directional enrichment approach, method 3 represents the ranked genes approach, and method 4 represents the pathway-based approach. *Two compounds galantamine and memantine are FDA approved to treat Alzheimer's disease and were identified as significant results prior to multiple corrections. **velnacrine is an FDA approved drug for cognitive impairment and was identified as the 8th most significant result when querying the AD gene expression signature in the CMap database using method 4. Compounds highlighted in red are FDA approved to treat AD or cognitive impairment.*

4.4 Discussion

This chapter re-processed the raw CMap data using appropriate up-to-date data processing techniques and developed four different methods to query the reprocessed CMap database for drug repositioning based on gene expression profiling. The four methods go beyond the constraints of the current CMap algorithm by not limiting the input gene expression signature by 500 genes and by considering both the up-regulated and down-regulated genes in a disease signature to identify more relevant candidate compounds. The four methods are bi-directional enrichment (method 1), anti-correlation (method 2), ranked genes-based (method 3) and a pathway-based approach (method 4).

4.4.1 The drug-drug validation suggests all four drug repositioning methods are valid

The drug-drug validation analysis used five random compounds from the CMap database, reversed their expression profile and used them as "dummy" disease gene expression signatures. All four methods successfully identified all five compounds in the CMap database as the most significant candidate compound, suggesting each method can identify candidate compounds which has a reverse gene expression profile to the input query signature. However, the number of significant candidate compounds identified by each method varied, with no apparent relation between the remaining candidate compounds across methods. In addition, a positive correlation was observed between the input DEG length and the number of candidate compounds identified by each method, suggesting disease signatures with more DEG's may identify more candidate compounds. However, during the disease-drug validation process, this phenomenon was not observed, suggesting many compounds within the CMap database may have similar expression profiles.

4.4.2 Method validation fails to identify the most effective drug repositioning technique

To evaluate the effectiveness of each method's capability of identifying relevant candidate compounds in relation to the input profile, three disease gene expression signatures were queried in the CMap database. The three disease gene expression profiles represented tuberculosis, breast cancer and prostate cancer. Samples were acquired prior to treatment, and current FDA approved treatment for these diseases were available in the CMap database. The underlying hypothesis of this approach assumed the gene expression profile of each disease dataset, after processing, would primarily be the result of the disease. Therefore, the best drug-repositioning method would correctly and consistently identify FDA approved compounds that are routinely used to treat each of the three diseases.

The only method to correctly identify at least one candidate compounds for all three diseases was the anti-correlation approach, however, on average across the three diseases, this method identified 606 candidate compounds of which four were FDA approved to treat the disease. When compared to the remaining three methods and the fact this method identifies approximately half the compounds in the

CMap database as a significant result, this method has the lowest specificity, and as such, can be deemed the worst approach from the four methods for drug repositioning. Therefore, this method was discarded from further use.

The bi-directional enrichment approach correctly identified the treatments for prostate cancer and breast cancer but failed to identify the correct candidate compounds for tuberculosis. In contrast, the pathway-based approach identified the correct treatment option for tuberculosis but failed for the cancer datasets. This may suggest, different methods may be suitable for different diseases and/or tissue sources.

The ranked gene-based method was very reserved in comparison to the remaining three methods, identifying a total of 7 candidate compounds for all three diseases, none of which were in the TTD as a treatment option for the three diseases. The number of candidate compounds identified by this method contradicted the drug-drug validation results, where a positive correlation was observed between the number of candidate compounds identified and the number of DEG's in the disease signature. This can suggest that this method may possibly be less prone to false positives. When querying the tuberculosis disease signature, only one significant result was identified; cefsulodin – an FDA approved treatment for bacterial infection. There is no relation of cefsulodin and tuberculosis in the TTD database; however, further literature investigation identifies this compound to be associated with tuberculosis (He *et al.*, 2015), although it is unknown if it can be used to treat the disease. If this compound is indeed a treatment option for tuberculosis, this ranked gene-based method would be the most sensitive method for drug repositioning. However, it cannot be disregarded that antibiotics are the most abundant drugs in the CMap database, and the result obtained is by chance.

The CMap experiments were performed on various cell lines; however, only the MCF7 cell line was analysed due to containing the highest number of instance expression profiles. Gene expression is tissue-specific (Aguet *et al.*, 2017) and drug-gene expression signatures observed in the MCF7 cell line may not reflect gene expression signatures of different tissues, such as those in the disease-drug validation (primary dendritic cells for tuberculosis, mammary gland for breast cancer and prostate epithelium for prostate cancer). Different methods were observed to work well with different disorders, which can be the consequence of different methods compensating for tissue-specific expression effects. However, it cannot be disregarded that significant results obtained from each method could be potential compounds to treat the disease or are false positives. The current results cannot conclude which approach is the best-suited method for drug repositioning the CMap based on gene expression profiling. Additional disease datasets for disease-drug validation that satisfy the inclusion criteria and additional experimental validation can be of benefit, albeit it's beyond the current scope of this thesis.

4.4.3 Candidate compound gabazine identified to treat AsymAD

Drug repositioning the AsymAD brain gene expression signature identified a number of candidate compounds. The Bi-directional enrichment method identified compound SR-95531, which according to TTD was not approved to treat any disease; however, additional literature investigation discovered SR-95531 is also referred to as gabazine (Behrens, van den Boom and Heinemann, 2007), which although is not FDA approved to treat any disease, is a $GABA_A$-receptor antagonist. Gamma-aminobutyric acid (GABA) is the principal inhibitory neurotransmitter in human CNS and most aspects of the GABA signalling system including the expression levels and functional characteristics of GABA receptors and transporters, are affected in AD (Kwakowsky *et al.*, 2018). The activity of GABA can be inhibited through three receptor families; $GABA_A$, $GABA_B$ and $GABA_C$ receptors (Li *et al.*, 2016). Reduced long-term potentiation (LTP) has been suggested to be linked to enhanced $GABA_A$ receptor-mediated inhibition, and several independent rodent studies have observed improved cognitive functions by long term $GABA_A$ antagonists (Yoshiike *et al.*, 2008). Gabazine also has a BBB permeability probability of 0.97 according to admetSAR, suggesting it is highly likely to penetrate the BBB, making it an ideal candidate compound for AD. To date, no literature was found to assess the effects of gabazine in AD, and further experimental investigation is warranted.

4.4.4 FDA approved treatments for AD and cognitive impairment may be effective in AsymAD

The ranked gene-based approach identified tacrine as the most significant candidate compound when repositioning the AsymAD gene expression disease signature in the CMap database. Tacrine is a cholinesterase inhibitor, and the first-ever FDA approved drug to treat AD which has been shown to improve cognition in mild to moderately impaired patients, although the treatment does not affect the course of the disease (Crismon, 1994). However, the drug was not widely adopted as periodic blood tests were required to monitor hepatotoxicity and many side-effects were reported including nausea, vomiting, dizziness, diarrhoea, seizures and loss of consciousness (Mehta, Adem and Sabbagh, 2012). Subsequently, tacrine has been replaced in many countries by safer alternative cholinesterase inhibitors galantamine, rivastigmine and donepezil to treat AD. Galantamine was also identified as a candidate compound (6^{th} most significant) with the ranked genes-based method; however, it did not pass multiple corrections. This suggests the discontinued AD treatment tacrine may be statistically and computationally more effective in reversing the gene expression signature of AsymAD patients than the current FDA-approved AD treatment galantamine.

A different drug repurposing approach based on biological pathway perturbations (method 4) identified thioperamide as a candidate compound to treat AsymAD, which is FDA approved to treat cognitive

impairment. A deficit in cholinergic neurotransmitters has been suggested to be the primary cause of cognitive decline in AD, with other neurotransmitter systems such as neuronal histamine suggested to further contribute to the development of AD. Therefore, targeting the histamine system for AD treatment is an obvious choice, although it has been largely neglected (Zlomuzica *et al.*, 2016). Drugs targeting the histamine system have been suggested to enhance the degradation of Aβ and NFT by promoting brain autophagy whilst stimulating neurogenesis to counteract memory-related cell loss (Zlomuzica *et al.*, 2016). However, human clinical trials in mild-to-moderate AD have failed to show specific histamine antagonists can improve cognition (Grove *et al.*, 2014), with future studies recommended to investigate novel histamine-related drugs (Zlomuzica *et al.*, 2016). Thioperamide (identified as a candidate compound for AsymAD) is a histamine antagonist and has not been previously associated with AD, and therefore is a prime candidate for further experimental investigations.

The candidate compounds discussed above (tacrine and thioperamide) have been identified in this thesis to treat AsymAD, which coincidently have already been FDA approved to treat AD and cognitive impairment respectively. It is important to note the FDA approved compounds available in the CMap database to treat AD (tacrine, galantamine and memantine) and cognitive impairment (thioperamide and velnacrine) were not identified as significant candidate compounds during the drug-drug or disease-drug validation processes (excluding the anti-correlation approach which was discarded due to increase of false positives), and therefore provides confidence the results obtained from repurposing the AsymAD disease signature may be biologically relevant. As these compounds were not identified by repurposing the AD gene expression signature as well, these compounds may be more effective in the asymptomatic stage of the disease. Nevertheless, it is important to remember these compounds are prescribed for AD symptomatic relief and do not modify the underlying pathology of the disease.

4.4.5 Anti-inflammatory drug epirizole identified to treat AsymAD

Inflammation response has been suggested to contribute to neurodegeneration in AD patients and has been suggested to be induced by Aβ and NFT interfering with a neuronal activity which activates inflammatory activities of microglia (Bolós, Perea and Avila, 2017). The potential of a nonsteroidal anti-inflammatory drug (NSAID) to treat AD was postulated when rheumatoid arthritis patients who were on NSAID were discovered to be at a lower risk of dementia (Martyn, 2003). Mouse models have also demonstrated NSAID such as ibuprofen have been found to reduce the production of β42 peptides (Weiner and Selkoe, 2002); however, AD clinical trials involving anti-inflammatory drugs have had contradictory results, with discrepancies across studies possibly explained due to different trials using different anti-inflammatory drugs. In addition, it has been suggested NSAID may be ineffective in AD once the disease has been established but may be beneficial in the presymptomatic phase of the disease (Martyn, 2003). A study investigated this hypothesis discovered NSAIDs have an adverse effect

in symptomatic AD, however, treating AsymAD patients reduced the incidence of AD, but only after 2-3 years of continuous administration (Breitner *et al.*, 2011).

Epirizole is an FDA approved NSAID and the second most significant candidate compound identified when repurposing the AsymAD gene expression signature using the ranked gene-based method in the CMap database. According to admetSAR, epirizole is highly predicted to penetrate the BBB, which is further supported by experimental evidence that demonstrated epirizole is able to pass the BBB and possess CNS activity (Dileep *et al.*, 2013). However, to date, no clinical trial has attempted to assess the effects of this compound in AD. The current results combined with literature review suggests epirizole may potentially have therapeutic effects in AsymAD, and further investigation is required.

4.4.6 An anti-bacterial compound identified to treat AsymAD and AD

Anisomycin was identified as the most significant compound to treat AsymAD subjects according to drug repositioning method 1. Anisomycin is an antibiotic that inhibits bacterial protein and DNA synthesis, and is predicted (probability 0.8) by admetSAR to be able to penetrate the BBB. A study investigated the effects of administration of anisomycin in an animal model of posttraumatic stress disorder and observed reduced anxiety-like and avoidant behaviour (Cohen *et al.*, 2006). These extreme behaviour responses are involved in memory consolidation; therefore, the study concluded the results suggests that anisomycin disrupts traumatic memory consolidation in rats. Although no literature association has been found between anisomycin and AD, the drug repositioning method in this study has identified a compound that has been suggested to be involved affecting memory, which is the hallmark symptom of AD, thus, warranting further investigation. 32

The most significant candidate compound identified for AD patients was paromomycin, an antibiotic used for the treatment of acute and chronic intestinal amebiasis. Although the CMap database contains the most anti-bacterial compounds (66/1146) and the fact anti-bacterial compounds were identified as candidate compounds during the disease-drug validation suggesting these compounds may represent false positives, the p-value associated with paromomycin result (p-value = 0) cannot be ignored. The literature on paromomycin is limited, with no known association with AD. Pathogenic components have been long suspected of playing an essential role in the onset and progression of AD. A recent study identified porphyromonas gingivalis (p.gingivalis), a pathogenic bacterium, in human AD brains and demonstrated mice infected with the same bacterium resulted in the release of toxic proteases known as gingipains which increased the production of $A\beta_{1-42}$, a component of amyloid plaques (Dominy *et al.*, 2019). Subsequently, the authors demonstrated a broad spectrum of antibiotics (did not test paromomycin) failed to inhibit the cytotoxic effects of p. gingivalis, however, a pre-designed gingipain

inhibitor reduced the bacterial brain load, blocked $A\beta_{1-42}$ productions, reduced neuroinflammation, and rescued neurons in the hippocampus.

Amyloid-β peptides have also been shown to protect against microbial infections in mice (Kumar *et al.*, 2016), providing further evidence of bacterial involvement in AD. Therefore, the antibiotic paromomycin that was identified when repurposing the AD gene expression signature may be of therapeutic benefit and as admetSAR predicts (probability = 0.87) the compound to be able to penetrate the BBB, paromomycin also requires further investigation.

4.4.7 Limitations

The four-drug repositioning methods developed to query the CMap database all have their strengths and weaknesses. The first method was the bi-directional enrichment approach which uses a hypergeometric test to assess the overlap of genes in the disease signature and every CMap compound. This method is a simple, quick and practical approach that is commonly used to associate gene lists to biological pathways through GSEA. However, this method ignores the magnitude of gene expression fold change and may, therefore, suggest candidate compounds for diseases based on "weak" relations. For instance, if gene A in a disease is significantly up-regulated 10-fold, and compound X significantly down-regulates gene A, but only by 0.1-fold, the bi-directional approach would match this compound to the disease signature, however, the compound may not be as effective at countering the effects of gene A. This issue could be addressed merely by adjusting the concentration of compound X to increase its effect on gene A, although this gets extremely complex and may not be effective when many genes are perturbed.

The second method was a simple anti-correlation approach where DEG's in the disease gene expression profile is correlated to the same genes in a CMap compound. This method assumes each compound in the CMap database can affect the gene expression of all genes, and it's the directional change of the gene that is important. Therefore, all genes in both the disease and compound expression profiles are assigned a fold change of either +1 or -1. However, the basis of this method is also a limitation as the method fails to consider if a gene is significantly affected by a compound (does not use p-values). Similar to method 1, the method ignores the magnitude of gene expression fold change and may, therefore, suggest candidate compounds for diseases based on "weak" relations.

The third approach used a ranked gene-based method, which incorporated the magnitude of fold change and the statistical significance of a gene being perturbed. This method addressed the flaws of the first two methods but is more computationally intensive. The fourth method used information across the perturbed genes to identify disrupted biological pathways. This method is free from gene-level similarities and focuses on the collective functionality of DEG's. Therefore, a candidate compound

does not need to directly counteract against the same genes affected in the disease gene expression, but instead the candidate compound needs to counteract against the biological effects of the DEG's in the disease. This pathway-based method uses SPIA to identify biological pathways as perturbed from a gene list; however, the SPIA method is limited to the number of biological pathways it queries. The SPIA method only contains information on 139 biological pathways when other biological pathway databases such as the ConsensusPathDB contains information on over 4000 biological features (Kamburov *et al.*, 2009). However, the SPIA incorporates topological information, including PPI to determine if a biological pathway is likely to be "inhibited" or "activated", which is vital to identify candidate compounds. This information is missing in other publicly available databases such as ConsensusPathDB. The largest repository identified with this information is the KEGG database (Ogata *et al.*, 1999), which contains information on over 300 biological pathways; however, this version of the database is only available at a commercial cost, which could not be financially supported by this thesis.

The CMap experiments were performed on various cell lines. However, only the MCF7 cell line was analysed as it contained the greatest number of unique compound expression profiles. Gene expression is tissue-specific (Aguet *et al.*, 2017) and drug-gene expression signatures observed in the MCF7 cell line may not reflect gene expression signatures of different tissues, such as the brain tissue. Therefore, this work purely identifies candidate compounds that need further experimental validation.

Finally, clinical phenotypes were somewhat limited for the AsymAD or AD subjects, including medication. Therefore, it is unknown if these subjects were on any medication that may have influenced gene expression signatures. However, as the AsymAD group had no clinical mention of dementia and were assumed to be clinically healthy until an autopsy, they would have been highly unlikely to be on medication relating to AD, and therefore the AsymAD gene expression profile most likely represented the disease alone. However, the AD group were all clinically and autopsy-confirmed AD and would have most likely been on AD medication, and therefore the AD gene expression profile obtained in this thesis most likely represented the combination of disease and effect of any medication. Nevertheless, as there is currently no pathological modifying treatment for AD, the AD gene expression signature may have captured gene expression representing hallmark AD pathology and the candidate compounds identified by repurposing the AD gene expression signature may still be of therapeutic value.

4.5 Conclusion

Drug repositioning has been demonstrated to be an effective and quick method for repurposing FDA approved compounds for new therapeutic targets. In this chapter, the CMap database was reprocessed using appropriate up-to-date processing techniques, and new sophisticated drug repositioning methods were developed and deployed to identify candidate compounds for AD. FDA approved

compounds to treat AD, and cognitive impairment was identified as candidate compounds for AsymAD, but not AD, suggesting these approved treatments may be more effective in the asymptomatic stage of the disease. Additional candidate compounds epirizole (anti-inflammatory compound) and paromomycin (antibiotic) were identified for AsymAD and AD respectively. These compounds warrant further experimental investigation to identify any interactions with the disease.

Chapter 5: Working Towards a Blood-Derived Gene Expression Biomarker Specific for Alzheimer's Disease

5.1 Background

Alzheimer's disease (AD) is a progressive neurodegenerative disorder affecting an estimated one in nine people over the age of 65 years of age, making it the most common form of dementia worldwide ('2018 Alzheimer's disease facts and figures includes a special report on the financial and personal benefits of early diagnosis', 2018). Current clinical diagnosis of the disease is primarily based on a time-consuming combination of physical, mental and neuropsychological examinations. With the rapid increase in the prevalence of the disease, there is a growing need for a more accessible, cost-effective and time-effective approach for early diagnosis and monitoring AD.

For research purposes, brain positron emission tomography (PET) scans and cerebral spinal fluid (CSF) have been used for disease identification. In particular, decreased Aβ and increased tau levels in CSF have been successfully used to distinguishing between AD, mild cognitive impairment (MCI) and cognitively healthy individuals with high accuracy. However, as a relatively invasive and costly procedure, it may not appeal to the majority of patients or be practical on a large-scale trial basis for screening the population (Han *et al.*, 2013; Lunnon *et al.*, 2013; Henriksen *et al.*, 2014). Peripheral blood-derived biomarkers could potentially be a solution to this problem.

Blood is a complex mixture of fluid and multiple cellular compartments that are consistently changing in protein, lipid, RNA and other biochemical entity concentrations (Thambisetty and Lovestone, 2010), which may be useful for AD diagnosis. Nakamura et al. successfully used $APP_{669–711}$/ $Aβ_{1–42}$ and $Aβ_{1–40}$/$Aβ_{1–42}$ ratios, and their composites, to predict individual brain Aβ load when compared to Aβ-PET imaging (Nakamura *et al.*, 2018). However, the test predicts Aβ, which is also found in other brain disorders such as FTD and therefore, the test requires AD specificity evaluation. Another study reviewed 163 candidate blood-derived proteins as potential AD biomarkers from 21 separate studies (Kiddle *et al.*, 2014). The overlap of biomarkers between studies was limited, with only four biomarkers α-1-antitrypsin, α-2-macroglobulin, apolipoprotein E and complement C3 found to replicate in five independent cohorts. However, a follow-on study discovered these biomarkers were not specific to AD, and were also discovered to be associated with other brain disorders including Parkinson's Disease (PD) and Schizophrenia (SCZ) (Chiam *et al.*, 2014), suggesting the need to consider other neurological and related disorders in study designs to enable the discovery of biomarkers specific to AD.

Similarly, several studies have attempted to exploit blood transcriptomic measurements for AD biomarker discovery. Initial research was limited to the analysis of single differentially expressed genes

(DEG) as a means to distinguish AD from cognitively healthy individuals (Rye *et al.*, 2011; Han *et al.*, 2013). However, the limited overlap and reproducibility of DEG from independent cohorts suggests this method alone is not reliable enough (Han *et al.*, 2013). A solution to this problem would be to use information across all genes simultaneously through machine learning algorithms to identify combinations of gene expression changes that may represent a biomarker for AD. This technique has been employed in multiple studies, which have demonstrated the ability to differentiate AD from non-AD subjects (Fehlbaum-Beurdeley *et al.*, 2010; Booij *et al.*, 2011; Lunnon *et al.*, 2013; Roed *et al.*, 2013; Voyle *et al.*, 2016). However, small sample size and lack of independent validation datasets most likely led to overfitting. The decrease in costs associated with microarray technologies led a study developing an AD classification model based on a larger training set of 110 AD and 107 controls and validating in a larger independent cohort of 118 AD and 118 controls. The model achieved 56% sensitivity, 74.6% specificity, and an accuracy of 66%, which equated to 69.1% Positive Predictive Power (PPV) and 63% Negative Predictive Power (NPV) (Voyle *et al.*, 2016). This was one of the first studies to demonstrate validation in an independent cohort; however, the classification model still lacked the 90% predictive power desired from a clinical diagnostic test (Huynh and Mohan, 2017).

Previous studies have demonstrated the potential use of blood transcriptomic levels to differentiate between AD and cognitively healthy individuals; however, they are yet to be precise enough for clinical utility and/or are yet to be extensively evaluated on specificity by assessing model performance in a heterogeneous ageing population of multiple diseases. This validation process is critical to determine whether the classification model is indeed disease-specific, a general indication of ill health, or an overfit.

This study developed a novel XGBoost classification model trained on blood transcriptomic profiling from AD, related mental disorders (Parkinson's disease [PD], Multiple Sclerosis [MS], Amyotrophic Lateral Sclerosis [ALS], Bipolar Disorder [BD], Schizophrenia [SCZ]), age-related disorders (Coronary Artery Disease [CD], Rheumatoid Arthritis [RA], Chronic Obstructive Pulmonary Disease [COPD]), and cognitively healthy subjects to differentiate AD from diseased and otherwise normal subjects. The classification model was developed with clinical utility in mind, with each dataset processed and transformed independently and evaluated in an independent ageing heterogeneous population (testing set) consisting of similar diseases as the training set.

5.2 Methods

5.2.1 Data acquisition

Microarray gene expression studies were sourced from publicly available repositories GEO (https://www.ncbi.nlm.nih.gov/geo/) and ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) in May

2018. Study inclusion criteria were; 1) microarray gene expression profiling must be performed on an age-related or neurological disorder, 2) RNA was extracted from whole blood or a component of blood, 3) study must contain at least ten subjects, and 4) data was generated on either the Illumina or Affymetrix microarray platform using a BeadArray containing at least 20,000 probes. The microarray platform was restricted to Affymetrix and Illumina only, as replication between the two platforms is generally very high (Barnes *et al.*, 2005), and BeadArrays restricted to a minimum of 20,000 probes to maximise the overlap of genes across studies, while also optimising the number studies available for inclusion.

5.2.2 Data processing

The data processing pipeline was designed with reproducibility and clinical utility in mind. New subjects could be independently processed and predicted through the same classification model without using any prior knowledge on gene expression variation of the data used to develop the classification model and without making any alteration to the classification model itself. All data processing was undertaken in RStudio (version 1.1.447) using R (version 3.4.4). Microarray gene expression studies were acquired from public repositories using the R packages "GEOquery" (version 2.46.15) and "ArrayExpress" (version 1.38.0). For longitudinal studies involving treatment effect, placebo subjects or initial gene expression profiling from baseline subjects before treatment were used. Studies consisting of multiple disorders were separated by disease into datasets consisting of diseased subjects and corresponding healthy controls if available.

Raw gene expression data generated on the Affymetrix platform were "mas5" background corrected using R package "affy" (version 1.42.3), log2 transformed and then Robust Spline Normalised (RSN) using R package "lumi" (version 2.16.0). Datasets generated on the Illumina platform were available in either a "raw format" containing summary probes and control intensities with corresponding p-values or a "processed format" where data had already been processed and consisted of a subset of probes and samples deemed suitable by corresponding study authors. When acquiring studies, preference was given to "raw format" data where possible, and when available, was "normexp" background corrected, log2 transformed, and quantile normalised using the "limma" R package (version 3.20.9).

Sex was then predicted using the R package "massiR" (version 1.0.1) and subjects with discrepancies between predicted and recorded sex removed from further analysis. Next, within each gender and disease diagnosis group of a dataset, probes above the "X" percentile of the log2 expression scale in over 80% of the samples were deemed "reliably detected". To account for the variation of redundant probes across different BeadArrays, the "X" percentile threshold value was manually adjusted until a variety of robust literature defined house-keeping genes were correctly defined as expressed or

unexpressed in their corresponding gender groups (Chang *et al.*, 2011). Any probe labelled as "reliably detected" in any group (based on gender and diagnosis) was taken forward for further analysis from all samples within that dataset. This substantially eliminates noise (Lazar *et al.*, 2013) and ensures disease and gender-specific signatures are captured within each dataset.

Next, to ensure homogeneity within biological groups, outlying samples were iteratively identified and removed using the fundamental network concepts described in (Oldham, Langfelder and Horvath, 2012). Finally, to enable cross-platform probes to be comparable, platform-specific probe identifiers were annotated to their corresponding universal Entrez gene identifiers using the appropriate BeadArray R annotation files; "hgu133plus2.db", "hgu133a.db", "hugene10sttranscriptcluster.db", "illuminaHumanv4.db" and "illuminaHumanv3.db".

5.2.3 Cross-platform normalisation

To enable transcriptomic information between datasets to be directly comparable, a rescaling technique, the YuGene transform, was applied to each dataset independently. YuGene assigns modified cumulative proportion value to each measurement, without losing essential underlying information on data distributions and allows the transformation of independent studies and individual samples (Lê Cao *et al.*, 2014). This allows new data to be added without global renormalisation and enables the training and testing data to be independently rescaled. Common probes across all processed datasets that contained both female and male subjects were extracted from each dataset and independently rescaled using the R package YuGene (version 1.1.5). YuGene transformation assigns a value between 0 and 1 to each gene, where 1 is highly expressed. As samples originated from publicly available datasets, potential duplicate samples may exist in this study. To address this issue, correlation analysis was performed on all samples using the common probes. Any sample with a Pearson's correlation coefficient equal to 1 was suggested to be a duplicate sample and would be removed from further analysis.

5.2.4 Training Set and Testing Set assignment

Multiple datasets from the same disease were available, allowing entire datasets to be assigned to either the "Training Set" for classification model development or "Testing Set" for independent external validation purposes. Larger datasets from the same disease were prioritised to the training set, allowing the machine learning algorithm to learn in a larger discovery set.

Individual subjects within the training and testing set were assigned a "0" class if the individual was AD or "1" if the individual was non-AD (includes healthy controls and non-AD diseased subjects). Grouping the non-AD subjects into a single class effectively mimics a large heterogeneous ageing population

where subjects may have a related mental disorder, neurodegenerative disease, age-related disease or are considered relatively healthy.

5.2.5 Classification model development

Two classification models were created. (i) The first was developed using only the AD and associated control datasets available in the training set and is referred to as the "AD vs healthy control" classification model. (ii) The second classification model was developed using all datasets and diseases available in the training set (AD, non-AD disease and all controls) and is referred to as the "AD vs mixed control" classification model. The second approach aimed to develop a classification model that may be more specific in identifying AD than the typical "AD vs healthy control" classification model.

Classification models were built using a powerful tree boosting algorithm, XGBoost, which in 2015 was used in every winning team in the top 10 of the Data Mining and Knowledge Discovery competition for a wide range of machine learning problems (Chen and Guestrin, 2016) and was suggested to be one of the most sophisticated methods at the time of this work (Patil, Aghav and Sareen, no date; Dhaliwal *et al.*, 2018). Furthermore, the tree learning algorithm uses parallel and distributed computing and is approximately 10 times faster than existing methods and allows many hyperparameters to be tuned to reduce the chance of overfitting (Dhaliwal *et al.*, 2018).

First, the training sets were balanced by up-sampling the minority class with replacement to match the number of samples in the majority class. As the second classification model ("AD vs mixed control") consisted of multiple diseases and complementary controls in the training set, all the control samples were assumed to be healthy and were therefore pooled. Then, within each disease classes of the mixed control group were up-sampled to match the total number of samples in the pooled control group. The AD samples were then up-sampled to match the total number of samples in the mixed control group. This process would ensure that all diseases in the mixed control group had the same probability to be used during the model development process.

Next, the R package "xgboost" (version 0.6.4.1) was used to create optimised models. Default tuning parameters were set to **eta**=0.3, **max_depth**=6, **gamma**=0, **min_child_weight**=1, **subsample**=1, **colsample_bytree**=1, **objective**="binary:logistic", **nrounds**=5000, **early_stopping_rounds parameters**=20 and **eval_metric=**"logloss". The "**seed**" was randomly assigned "222" throughout the model developmental stages for reproducibility purposes. The initial model was built and internally evaluated using 10-fold cross-validation with stratification which calculates a test logloss mean at each **nrounds** iteration, stopping if an improvement to the test logloss means is achieved in the last 20 iterations. The **nrounds** iteration that achieved the optimal test logloss mean was used to build the initial classification model, reducing the chance for an "overfit" model.

During the internal cross-validation process, each feature (gene) was assigned an importance value ("variable importance feature") based on how well it contributed to the correct prediction of individuals in the training set. The higher the variable importance value for a gene, the more useful that gene was in distinguishing AD subjects from non-AD individuals. The genes contributing to the initial XGBoost model were each assigned a variable importance value. The least two variable important features were then iteratively removed, classification models re-built, and logloss performance measures re-evaluated. This process was repeated through all available baseline features, with the minimum logloss from all iterations used to determine the most predictive genes. This process is referred to as "recursive feature elimination" and has been shown to improve classification model performance and reduce model complexity by removing weak and non-predictive features (Guyon *et al.*, 2013).

Following identification of the most predictive genes, the classification model was further refined by iteratively tuning through the following hyperparameter values: **max_depth** (2:20, 1), **min_child_weight** (1:10, 1), **gamma** (0:10, 1), **subsample** (0.5:1, 0.1), **colsample_bytree** (0.5:1, 0.1), **alpha** (0:1, 0.1),), **lambda** (0:1, 0.1), and **eta** (0.01:0.2, 0.01), whilst performing a 10-fold cross-validation with stratification and evaluating the test logloss mean to select the optimum hyperparameters.

5.2.6 Classification model evaluation

The classification models were validated on the independent unseen testing set, predicting the diagnosis of all subjects as a probability ranging from 0 to 1, where AD ≤ 0.5 > non-AD. The prediction accuracy, sensitivity, specificity, PPV and NPV were calculated to evaluate the overall classification model's performance. To assess the sensitivity and specificity of the classifiers, ROC curves and AUC scores were generated using the R package "ROCR" (version 1.07) with the following recommended diagnostic interpretations used: "excellent" (AUC = 0.9-1.0), "very good" (AUC = 0.8-0.9), "good" (AUC = 0.7-0.8), "sufficient" (AUC = 0.6-0.7), "bad" (AUC = 0.5-0.6) , and "test not useful" when AUC value is <0.5 (Šimundić, 2009).

The clinical utility metrics were calculated to evaluate the clinical utility of the classification models. The positive Clinical Utility Index (CUI +) was calculated as PPV * (sensitivity/100) and the negative Clinical Utility Index (CUI -) calculated as NPV * (sensitivity/100). The Clinical Utility Index (CUI) essentially corrects the PPV and NPV values for the occurrence of that test in each respective population and scores can be converted into qualitative grades as recommended: "excellent utility" (CUI >= 0.81), "good utility" (CUI >=0.64) and "satisfactory utility" (CUI >=0.49) and "poor utility" (CUI < 0.49) (Mitchell, 2012). An overview of the classification model development and evaluation process is provided in Figure 5.1

5.2.7 The biological importance of predictive features

The final classification model taken forward for external validation contains a list of ranked genes which collectively differentiate AD from non-AD subjects. These genes have been derived from multiple disorders and may be involved with biological processes, which was therefore assessed through Gene Set Enrichment Analysis (GSEA). The predictive genes were analysed using an Over-Representation Analysis (ORA) implemented through the ConsensusPathDB (http://cpdb.molgen.mpg.de) web-based platform (version 33) (Kamburov *et al.*, 2009) in November 2018. For GSEA analysis, a minimum overlap of the query signature and database was set as 2.

5.2.8 Data availability

The data used in this study were all publicly available with accession details provided in Table 5.1. All analysis scripts used in this study are available at https://doi.org/10.5281/zenodo.3371459.

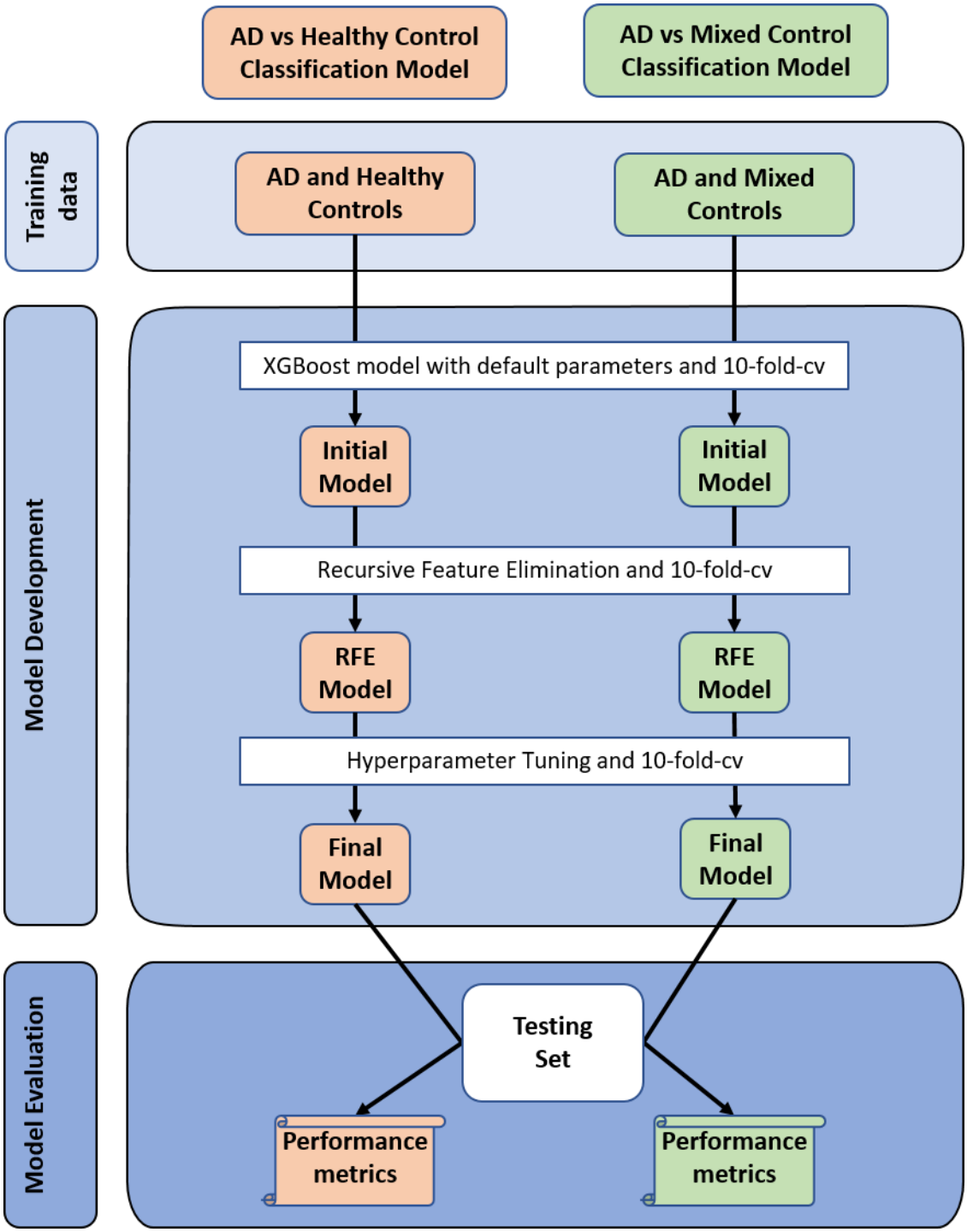


Figure 5.1: Overview of Study Design. "Logloss" metric was used throughout the model developmental stage to identify optimal features and hyperparameters. Abbreviations: cv = cross-validation and RFE = Recursive Feature Elimination.

5.3 Results

5.3.1 Summary of data processing

Twenty-one publicly available studies were identified, acquired and processed. Separating studies by disease status resulted in 22 datasets, which consisted of 3 AD, 3 MS, 3 SCZ, 3 CD, 3 RA, 2 COPD, 2 BD, 2 PD and 1 ALS orientated dataset. Only 15 datasets contained both diseased and complementary control subjects, while the remaining 7 contained only diseased subjects. An overview of the demographics of each dataset is illustrated in Table 5.1

Independently processing the 22 datasets resulted in a total of 2740 samples after Quality Control (QC), of which 287 samples were AD. Since 11 different BeadArrays had been used to expression profile the 9 different diseases, and as 7 datasets were only available in a "processed format" (GSE63060, GSE63061, E-GEOD-41890, GSE23848, E-GEOD74143, E-GEOD-54629 and E-GEOD-42296), each dataset varied in the number of "reliably detected" genes after QC (detailed in Table 5.1). Initially, an overlap of the common measurable probes across all 22 datasets which were also deemed "reliably detected" in any one of the datasets was compiled, resulting in 7452 genes. However, following the independent transformation of each dataset, platform and BeadArray-specific batch effects were observed (**Figure** 5.2a-b). This can be largely explained by different platforms having different probe designs to target different transcripts of the same gene, leading to significant discrepancies and even absence in the measurement of the same gene by different platforms (Barnes *et al.*, 2005). Therefore, to address this platform and BeadArray-specific batch effect, 1681 common "reliably detected" genes across all datasets that contained both male and female subjects (20 datasets) were extracted from each dataset and independently YuGene transformed. Essentially, these 1681 genes are expressed at a level deemed "reliably detected" in all 11 different BeadArrays. The distribution of the 1681 genes in each subject is shown in **Figure** 5.2c-d and can be seen to be more evenly distributed across the 2740 subjects than **Figure** 5.2a-b, a characteristic desired for the machine learning algorithms. Correlation analysis was then performed on all samples, which suggested all samples were highly correlated, with the maximum per sample correlation coefficients ranging from 0.86-0.99. No sample was deemed to be a duplicate, and therefore, no additional sample was removed following QC.

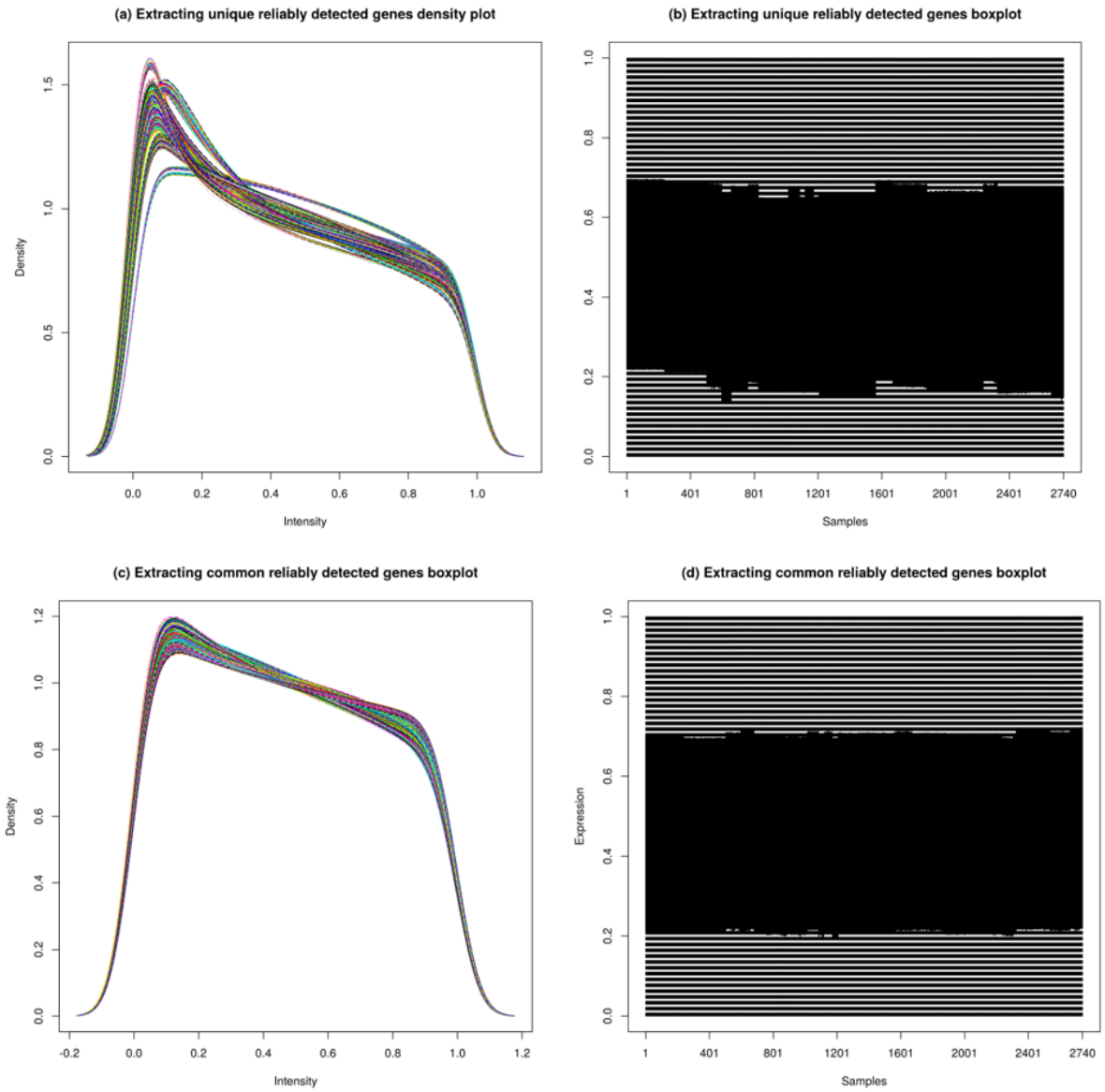


Figure 5.2: Shows the distribution of gene expression across all subjects. The distribution of gene expression across all individuals is expected to be similar when no batch effects exist. a) and c) are density plots, where each line represents the expression density of an individual while b) and d) represent the same data as boxplots respectively. a) and b) demonstrates batch effects caused by specific platform and BeadArrays when extracting 7452 unique genes deemed "reliably detected" in any of the 22 datasets. The shallow "n" curve in density plot a) are dominantly Illumina generated data, suggesting Illumina generated data have less lowly expressed genes in comparison to Affymetrix generated data. In contrast, c) and d) reveals a more evenly distributed gene expression profile across all subjects when extracting the 1681 common "reliably detected" genes, and therefore these genes were used for classification model development.

Table 5.1: Dataset Demographics

Disorder	Study ID (associated publication)	Platform	BeadArray	Tissue Source	Demographics Before QC			Samples Removed During QC			Demographics After QC				Training and Testing set Assignment
					No. Probes	Case Sex (M/F)	Control Sex (M/F)	No. Samples	No. Gender Mismatches	No. Outlying Sample	No. Probes	Case Sex (M/F)	Control Sex (M/F)	No. Samples	
Alzheimer's Disease	GSE63060 (Sood *et al.*, 2015)	I	HT-12 v3.0	WB	38323	46/99	42/62	249	2	10	5364	45/93	40/59	237	Training
	GSE63061 (Sood *et al.*, 2015)	I	HT-12 v4.0	WB	32049	51/81	55/87	274	5	4	5241	48/79	54/84	265	Testing
	E-GEOD-6613 (Scherzer *et al.*, 2006)	A	HG U133A	WB	22283	8/15	11/11	45	0	1	4184	8/14	11/11	44	Training
Parkinson's Disease	E-GEOD-6613 (Scherzer *et al.*, 2006)	A	HG U133A	WB	22283	38/12	0/0	50	0	0	3674	38/12	0/0	50	Training
	E-GEOD-72267 (Calligaris *et al.*, 2015)	A	HG U133A 2.0	PBMC	22277	23/17	8/11	59	0	0	8742	23/17	8/11	59	Testing
Multiple Sclerosis	GSE24427 (Goertsches *et al.*, 2010)	A	HG U133A	WB	22283	9/16	0/0	25	0	0	6633	9/16	0/0	25	Testing
	E-GEOD-16214 (De Jager *et al.*, 2009)	A	HG U133 plus 2.0	PBMC	54675	11/71	0/0	82	0	3	8098	11/68	0/0	79	Training
	E-GEOD-41890 (Irizar *et al.*, 2014)	A	Exon 1.0 ST	PBMC	33297	20/24	12/12	68	0	1	8157	19/24	12/12	67	Training
Schizophrenia	GSE38484 (de Jong *et al.*, 2012)	I	HT-12 v3.0	WB	48743	76/30	42/54	202	9	5	6700	69/28	39/52	188	Training
	E-GEOD-27383 (van Beveren *et al.*, 2012)	A	HG U133 plus 2.0	WB	54675	43/0	29/0	72	0	1	11297	42/0	29/0	71	Testing
	GSE38481 (de Jong *et al.*, 2012)	I	Human-6 v3	WB	24526	4/11	16/6	37	2	1	8106	11/3	15/5	34	Testing
Bipolar Disorder	E-GEOD-46449 (Clelland *et al.*, 2013)	A	HG U133 plus 2.0	L	54675	28/0	25/0	53	0	0	9882	28/0	25/0	53	Training
	GSE23848 (Beech *et al.*, 2010)	I	Human-6 v2	WB	48701	6/14	5/10	35	0	0	7211	6/14	5/10	35	Testing
Cardiovascular Disease	E-GEOD-46097 (Ellsworth *et al.*, 2014)	A	HG U133A 2.0	PBMC	22277	102/36	60/180	378	0	24	7676	94/36	57/167	354	Training
	GSE59867 (Maciejak *et al.*, 2015)	A	Exon 1.0 ST	WB	33297	85/26	0/0	111	0	3	7936	82/26	0/0	108	Testing
	E-GEOD-12288 (Sinnaeve *et al.*, 2009)	A	HG U113A	WB	22283	88/22	84/28	222	0	8	4815	83/22	82/27	214	Training
Rheumatoid Arthritis	E-GEOD-74143 (Walsh *et al.*, 2016)	A	HT HG U113 plus	WB	54715	81/296	0/0	377	1	23	8112	80/273	0/0	353	Training

	E-GEOD-54629 (Sellam *et al.*, 2014)	A	Exon 1.0 ST	WB	33297	11/58	0/0	69	0	0	11931	11/58	0/0	69	Testing
	E-GEOD-42296 (Mesko *et al.*, 2013)	A	Exon 1.0 ST	PBMC	33297	4/15	0/0	19	0	0	10417	4/15	0/0	19	Testing
Chronic Obstructive Pulmonary Disease	E-GEOD-54837 (Singh *et al.*, 2014)	A	HG U133 plus 2.0	WB	54675	91/45	57/33	226	0	16	5531	83/44	52/31	210	Training
	E-GEOD-42057 (Bahr *et al.*, 2013)	A	HG U133 plus 2.0	WB	54675	52/42	22/20	136	3	4	6445	49/39	21/20	129	Testing
ALS	E-TABM-940	A	HG U133 plus 2.0	WB	54675	27/26	18/19	90	3	10	10442	27/25	15/10	77	Training
Total						904/956	486/533	2879	25	114		870/906	465/49	2740	

Each study is accompanied by its corresponding publication (if available), where individual study design can be obtained. When possible, datasets were obtained in their raw format, except for GSE63060, GSE63061, E-GEOD-41890, GSE23848, E-GEOD74143, E-GEOD-54629 and E-GEOD-42296 which were only available in a processed form where dataset had already been background corrected, log2 transformed and normalised by techniques stated in corresponding publications. Multiple datasets from the same disease existed in this study. The dataset with the larger number of diseased subjects was prioritised into the training set for better discovery. Study ID's initiating with "GSE" and "E-GEOD" were obtained from GEO and ArrayExpress respectively. Abbreviations are as follows: I=Illumina, A=Affymetrix, WB=Whole Blood, PBMC=Peripheral Blood Mononuclear cell and L=Lymphocytes.

5.3.2 Training Set and Testing Set demographics

Multiple datasets from the same disease were obtained for this study, with the largest dataset from each disease assigned to the training set to improve discovery. However, three AD datasets were available, and the two largest datasets were generated on the Illumina platform, and the third on the Affymetrix platform. To address any subtle differences in gene expression which may still exist in the data due to platform differences, the largest Illumina AD and the Affymetrix AD datasets were both assigned to the Training Set.

Following dataset assignment, the training set consisted of 160 AD subjects and 1766 non-AD subjects, while the testing set consisted of 127 AD subjects and 687 Non-AD subjects. The Non-AD group in both the training and testing set consisted of subjects with either PD, MS, SCZ, BD, CD, RA, COPD or were relatively healthy. Only one ALS dataset suitable for this study was identified and was deemed too small to split into the training and testing set. Therefore, the ALS dataset was assigned to the training set, allowing the machine learning algorithm to learn multiple disease expression signatures, which could further aid in differentiating AD from Non-AD subjects. Samples in the training set were up-sampled to prevent biasing the majority classes during model development. This resulted in "AD vs healthy" classification model consisting of 160 AD samples and 160 complimentary healthy control samples, and the "AD vs mixed controls" being trained on 6318 AD samples and 6318 non-AD samples. The "AD vs mixed controls" training set contains significantly more samples as the pooled controls consisted of 702 samples; therefore, the remaining 8 disease classes were up-sampled to the same sample size which totalled 6318 samples. The AD samples were then up-sampled to 6318 to balance the training set. An overview of subjects in the training and testing set is provided in Table 5.2.

Table 5.2: Overview Training and Testing set subjects

Dataset	Training Set		Testing Set
	AD vs healthy control	AD vs mixed control	
Alzheimer's Disease	160*	6318 (160*)	127
Parkinson's Disease	0	702 (50)	40
Multiple Sclerosis	0	702 (122*)	25
Schizophrenia	0	702 (97*)	56*
Bipolar Disorder	0	702 (28)	20
Cardiovascular Disease	0	702 (235*)	108
Rheumatoid Arthritis	0	702 (353)	88*
Chronic Obstructive Pulmonary Disease	0	702 (127)	88
ALS	0	702 (52)	0
Pooled Controls	160 (127*)	702*	262

Entire datasets from each disease were assigned to either the "Training Set" for classification model development or the "Testing Set" for validation purposes. Datasets with a larger number of diseased subjects were prioritised into the training set to increase discovery. Two classification models were developed, the first was developed using only the 160 AD and associated 127 control samples, which were up-sampled to 160 to develop the "AD vs healthy control" classification model. The pooled controls in the "AD vs healthy control" training set originates only from AD datasets and can be regarded as cognitive healthy controls. The second classification model was developed using all datasets and diseases available in the training set and is referred to as the "AD vs mixed control" classification model, where samples in the minority classes are up-sampled to match the 702 samples in the pooled controls. The mixed control group totalled 6318 samples. Therefore, the AD group was up-sampled to 6318 to create a balanced training set. Sample numbers provided in brackets are before up-sampling. Sample numbers with an asterisk (*) indicates multiple datasets were available, and subject numbers shown are a sum across these datasets.

5.3.3 "AD vs healthy control" classification model development and performance

The "AD vs healthy control" classification model was developed using only the two AD datasets (GSE63060 and E-GEOD-6613) available in the training set, which after up-sampling consisted of 160 AD and 160 cognitive healthy controls. The model was initially built using default parameters which selected 126 predictive features from the available 1681 genes, resulting in a cross-validation test logloss mean of 0.47 (0.21 SD). Further refinement of the model identified 74 predictive genes and the optimum hyperparameters as **eta**=0.12, **max_depth**=10, **gamma**=0, **min_child_weight**=1, **subsample**=1, **colsample_bytree**=1, **alpha**=0, **lambda**= 1 and **nrounds** =63, which improved the test logloss mean to 0.27 (0.1 SD).

The "AD vs healthy control" classification model was validated on the independent testing set and achieved a sensitivity of 58.0%, specificity of 30.0% and a balanced accuracy of 44.3% (additional classification performance metrics are provided in Table 5.3). The probability predictions of individual samples in the testing set is illustrated in Figure 5.3a, where misclassification can be observed in all

diseases and controls, demonstrating an increased false-positive rate and the inability of the classification model to confidently assign a positive (0) or negative (1) class to each subject.

A ROC curve was generated for the "AD vs healthy control" classification model performance (Figure 5.4), which demonstrates a low TP rate in comparison to random and the AUC score of 0.49 suggests this "test is not useful" as a diagnostic test. The clinical utility values (CUI +ve = 0.08, CUI -ve = 0.24) mirrors the AUC score interpretation, as the CUI values suggest the classification model is "poor" at detecting the presence and absence of AD and based on current validation results, has no real clinical utility.

5.3.4 "AD vs mixed control" classification model development and performance

The "AD vs mixed control" classification model was developed on the entire training set, which after up-sampling consisted of 6318 AD and 6318 non-AD subjects. The classification model was built using default parameters which selected 231 genes from the available 1681 genes as predictive features and resulted in cross-validation test logloss mean of 0.015 (0.009 SD). Further refinement of the model identified 28 predictive features, **eta**=0.08, **max_depth**=6, **gamma**=0, **min_child_weight**=1, **subsample**=1, **colsample_bytree**=1, **alpha**=0, **lambda**=0.9 and **nrounds**=139 as the optimum parameters which improved the cross-validation test logloss mean to 0.009 (0.005 SD).

The "AD vs mixed control" classification model was further validated on the testing set and achieving 46.5% sensitivity, 95.6% specificity, and balanced accuracy of 71.0% (additional classification performance metric are provided in Table 5.3).

Table 5.3: Classification model performance

	"AD vs healthy control" classification model	"AD vs mixed control" classification model
Sensitivity	58.3%	46.5%
Specificity	30.3%	95.6%
Balanced Accuracy	44.3%	71.0%
PPV	13.4%	66.3%
NPV	79.7%	90.6%
AUC	0.49	0.84
AUC Rating	Test not useful	Very good
CUI +ve	0.08	0.31
CUI +ve Rating	Poor	Poor
CUI -ve	0.24	0.87
CUI -ve Rating	Poor	Excellent

The table provides the performance measurements form validating the "AD vs Healthy Control" and the "AD vs Mixed Control" classification models on the same testing set.

The performance of this classification model improves on the typical "AD vs healthy control" classification model in all performance metrics, except for sensitivity, where a decrease in performance is observed from 58% to 46.5%. Nevertheless, due to the "AD vs mixed control" classification model predicting less false positives, an increase in PPV (66.3%) is observed when compared to the "AD vs healthy control" classification model (PPV = 13.4%). Furthermore, as illustrated in Figure 5.3b, the probability predictions for individuals in the testing set are more correctly and confidently predicted when compared to the typical "AD vs healthy control" classification model, only misclassifying 21 pooled controls (8% of total pooled controls), 2 CD (2% of CD subjects), and 7 SCZ (13% of SCZ subjects) as AD. The "AD vs mixed control" classification model ROC curve (Figure 5.4) achieves an improved AUC score of 0.84 which translates to a "very good" diagnostic test, however, the clinical utility values (CUI +ve = 0.31 and CUI -ve = 0.87) suggests this classification model is "poor" in the detection of AD but "excellent" to rule out "AD".

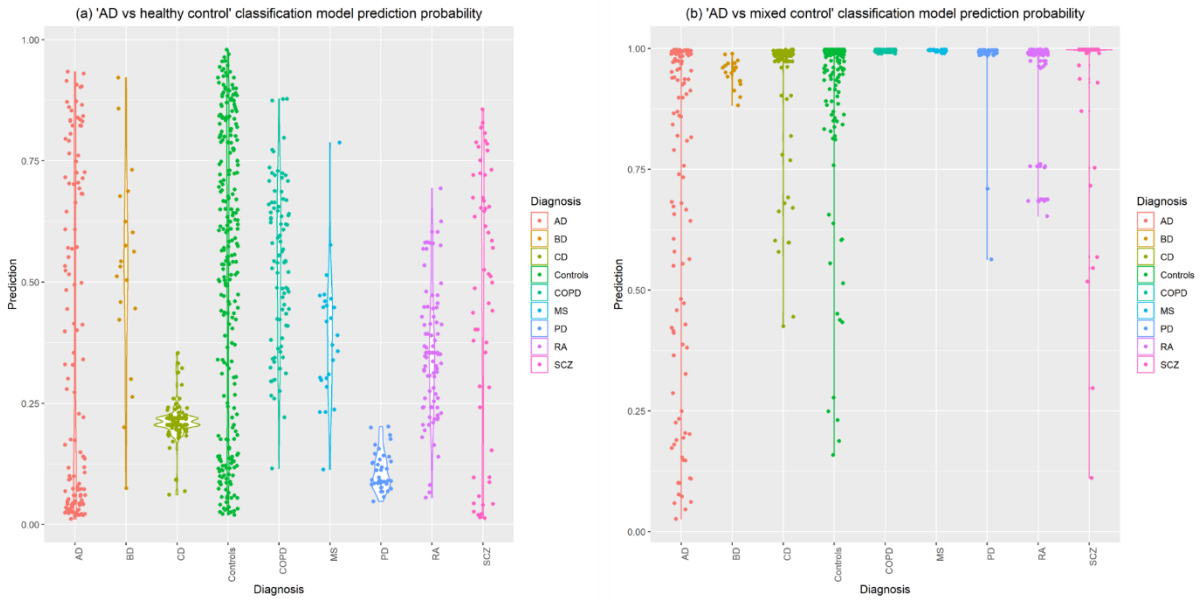


Figure 5.3: Shows the probability prediction of the testing set samples being AD (0) or non-AD (1). Subjects in the testing set represent a heterogeneous ageing population with subjects being clinically diagnosed with various mental-health related disorders, neurodegenerative diseases, age-related diseases or are relatively healthy. a) illustrates the confidence of the "AD vs healthy control" classification model distinguishing subjects in the testing set and b) illustrates the confidence of the "AD vs mixed control" classification model in predicting the same testing set. Controls represent pooled non-diseased subjects from all datasets. Diseases are abbreviated as follows; AD = Alzheimer's disease, BD = Bipolar disease, CD = Coronary Artery disease, COPD = Chronic Obstructive Pulmonary Disease, MS = Multiple Sclerosis, PD = Parkinson's Disease, RA = Rheumatoid Arthritis and SCZ = Schizophrenia.

5.3.5 "AD vs mixed control" classification model's predictive features

The variable importance was calculated for the 28 predictive genes (gene list provided in Supplementary Table 5.1) with the 20 most predictive genes illustrated in Figure 5.5. The gene LDHB

provides the greatest predictive value with a relative importance value of 0.7. GSEA performed on the 28 genes identified four biological processes significantly enriched; **WNT ligand biogenesis and trafficking** (p-value= 9.56e-4, q-value=0.04), **Alzheimer disease** (p-value= 3.70e-3, q-value=0.05),

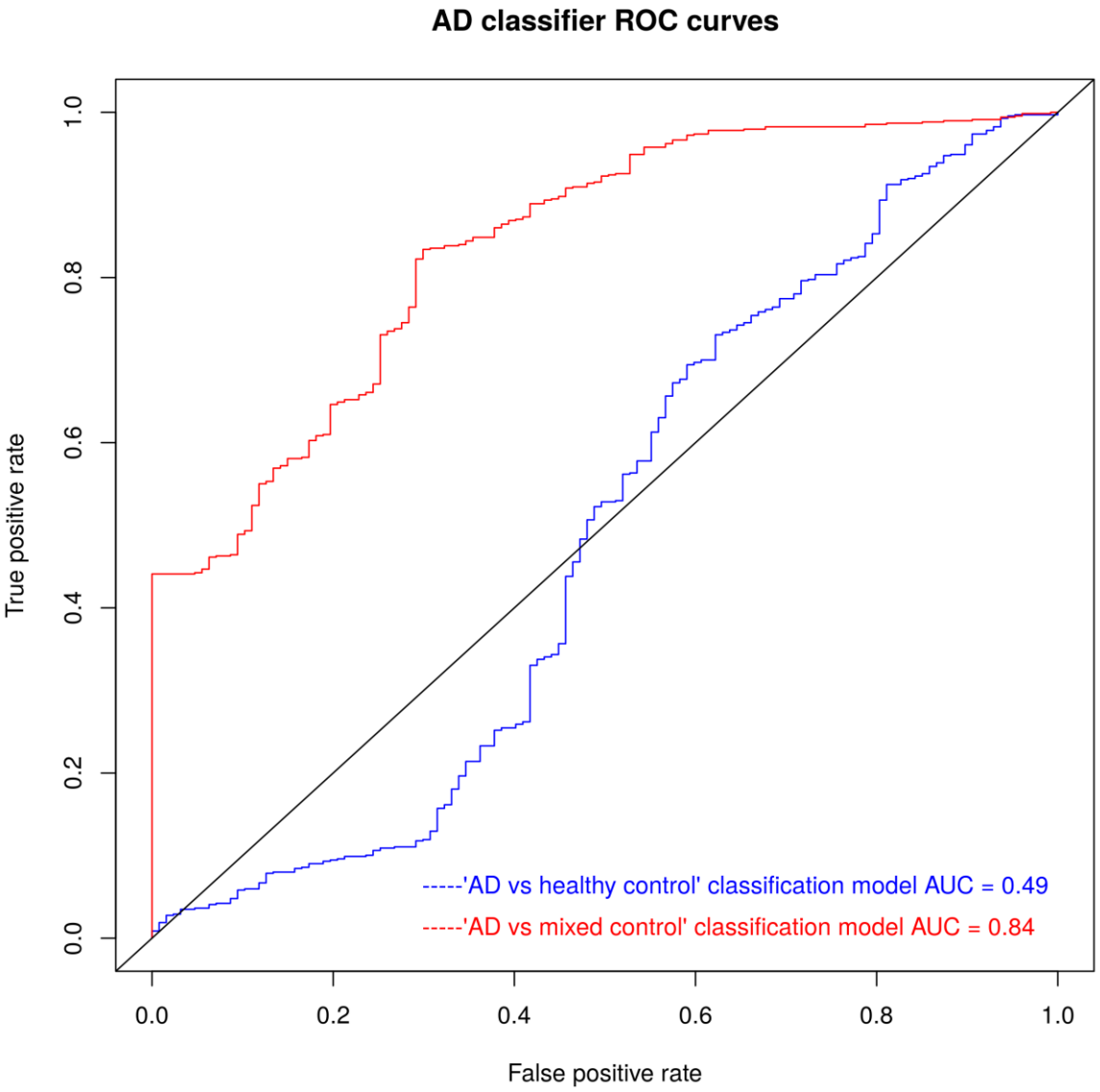


Figure 5.4: AD classification model ROC curves. The blue line represents the typical "AD vs healthy control" classification model's ROC curve which is trained using AD and complimentary healthy control subjects. The red line represents the "AD vs mixed control" classification model's ROC curve, which is trained using AD and non-AD diseased and healthy control subjects. The ROC curves were generated from the performance of both classification models in the testing set and demonstrate an improved TP and FP rate with the "AD vs mixed control" classification model, which achieved an improved AUC value of 0.84 when compared to the typical "AD vs healthy" classification approach which achieved an AUC score of 0.49

Herpes simplex infection (p-value= 4.61e-3, q-value=0.05), and **Huntington disease** (p-value= 5.19e-3, q-value=0.05). Additional information on gene overlap between the 28 predictive genes and biological pathways is provided in Supplementary Table 5.2.

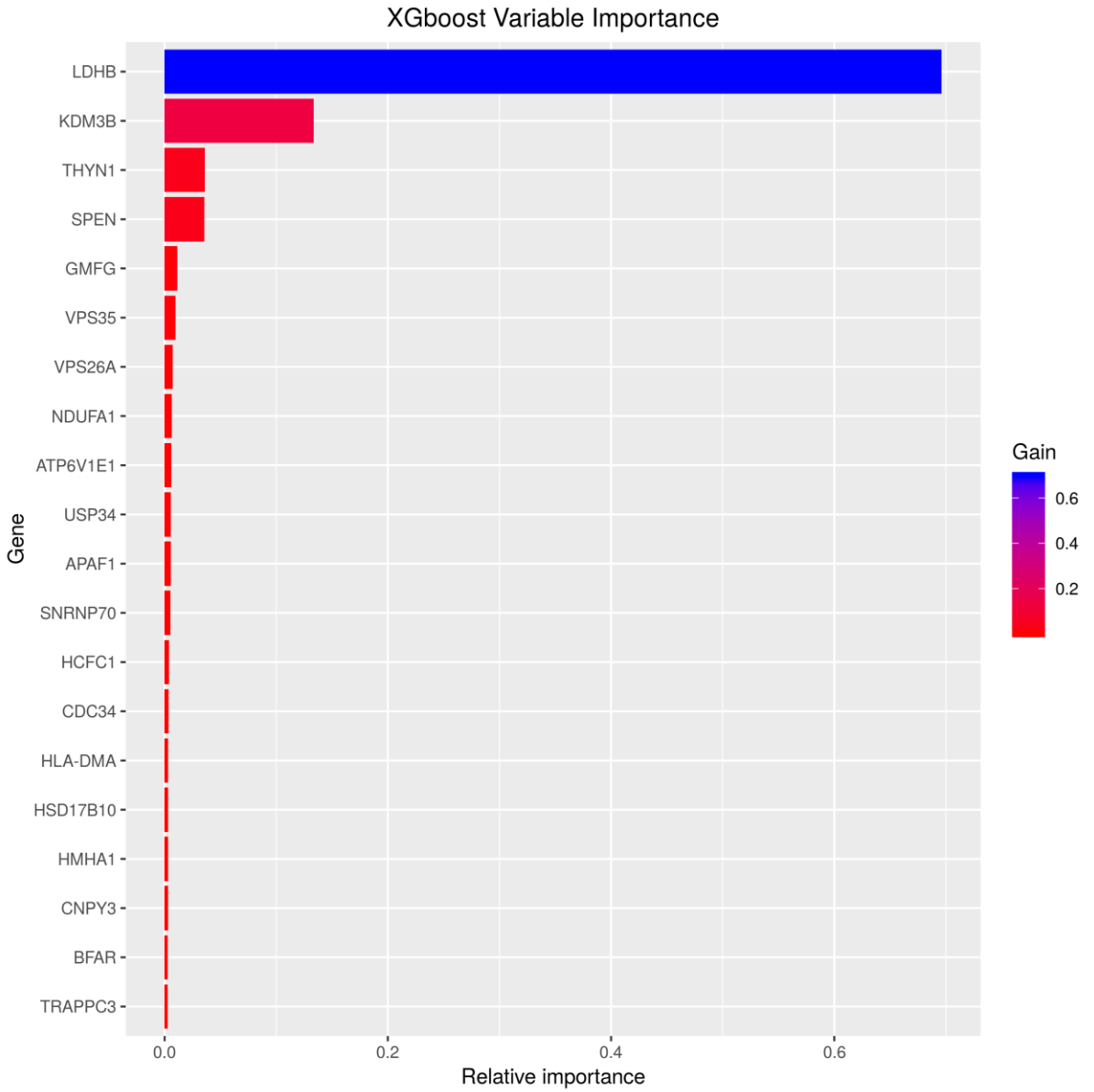


Figure 5.5: Relative Importance of the 20 most predictive genes for the "AD vs mixed control" classification model

5.4 Discussion

Previous attempts to identify blood-derived gene expression profiling for AD diagnosis have relied on the typical approach of training machine learning algorithms on AD and cognitively healthy subjects only, inadvertently leading to classification models learning expression signatures that may be of general illness rather than being disease-specific. Validating such a classification model in a heterogeneous ageing population may fail to distinguish AD from similar mental health disorders, neurodegenerative diseases, age-related disorders and cognitively healthy individuals. To address this issue, this study developed an "AD vs mixed control" classification model based on a training set comprised of AD, PD, MS, BD, SCZ, CD, RA, COPD, ALS and a set of pooled healthy individuals totalling 1926 subjects. The individual classes within the mixed control group varied in sample size, with the

pooled controls representing the largest class consisting of 702 samples. Therefore, to avoid sampling bias during the classification model development, the individual classes in the mixed control group were each up-sampled with replacement to 702, which totalled 6318 samples. The AD group were then up-sampled to 6318, classification model developed, optimised and evaluated in an external independent cohort comprised of similar diseases and controls totalling 814 subjects.

5.4.1 The typical "AD vs healthy control" classification model performs poorly in a heterogeneous ageing population

The typical approach of developing a classification model trained on AD and complimentary cognitive healthy control subjects produced a model with a sensitivity of 58.3% in an independent cohort of 127 AD subjects. The performance of this model is slightly better than a previous attempt which attained a sensitivity of 56.8% when validated in an independent testing set of 118 AD subjects (Voyle *et al.*, 2016). However, when evaluating this typical AD classification model in a heterogeneous ageing population, a process often neglected in previous studies, a very low specificity of 30.3% was attained which equated to a low PPV of only 13.4%. PD is the second most common cause of dementia ('2018 Alzheimer's disease facts and figures includes a special report on the financial and personal benefits of early diagnosis', 2018) and was seen to be most misclassified as AD. However, since misclassification was observed in all groups including large portions of the controls, this classification model is most likely not capturing signals of AD, dementia or general illness, but is most likely a result of technical noise, individual study batch effects and overfitting. This is mirrored in the model's performance metrics which translates to a "poor" clinical utility in detecting the presence and absence of AD. Overall, the typical approach of AD classification model development failed to accurately distinguish AD subjects in a heterogeneous ageing population consisting of PD, MS, BD, SCZ, CD, RA, COPD, ALS and relatively healthy controls.

5.4.2 The "AD vs mixed control" classification model outperforms the typical "AD vs healthy control" classification model

In contrast, the "AD vs mixed control" classification model attained a validation PPV of 66.3% and NPV of 90.6% on the same testing set, which outperforms the validation PPV of 13.4% and NPV of 79.9% achieved by the "AD vs healthy control" classification model. However, this improvement was at the cost of sensitivity, which was reduced from 58.3% ("AD vs healthy control") to 46.4% ("AD vs mixed control"). Nevertheless, an overall increase in the clinical utility of the "AD vs mixed control" classification model was measured and according to the recommended CUI interpretations in (Mitchell, 2012), the model is "poor" in "ruling in" AD but "excellent" in "ruling out" AD.

The performance of the "AD vs mixed control" classification model can be suggested to be superior due to the increased number of samples in the training set. The "AD vs healthy control" classification model was developed using 160 AD samples while the "AD vs mixed controls" classification model was developed using 6318 AD samples. However, it is important to note the AD samples in both training sets originated from the same subjects, with AD sample numbers in the "AD vs mixed control" training set up-sampled to account for the variation of sample sizes across the individual classes in the mixed control group. Therefore, the increased performance achieved by the "AD vs mixed control" classification model is most likely the result of incorporating additional related neurological and age-related disorders into the classification model development process, which aided in the identification of a more AD-specific expression signature than the typical approach of using only AD and corresponding control samples. Although this improved the ability to distinguish AD from other related diseases and cognitively healthy controls, the sensitivity of the model was reduced and needs to be further enhanced for this type of research to be beneficial in the clinical setting.

5.4.3 Predictive features are enriched for Herpes Simplex Infection

The underlying replication of predictive genes across blood-based transcriptomic biomarker studies are inconsistent (Ein-Dor *et al.*, 2005; Haury, Gestraud and Vert, 2011; Venet, Dumont and Detours, 2011; Siavelis *et al.*, 2016a). Nevertheless, sets of genes within independent studies have been able to consistently distinguish AD from complimentary controls (Siavelis *et al.*, 2016a). Therefore, the predictive features in this study warrant further investigation to assess their biological relevance to AD. The "AD vs mixed control" classification model differentiates AD from other diseases and healthy controls using the relationship of 28 genes. GSEA identified "Herpes simplex infection" as one of the biological pathways being significantly enriched prior to multiple corrections, with an overlap of 3 genes (CDC34, HCFC1 and HLA-DMA). This suggests gene expression changes associated with this process are measurable in blood and may contribute towards identifying AD subjects. Pathogenic viral components have been long suspected of playing an essential role in the onset and progression of AD. A recent study identified common viral species in normal and ageing brains, with an increased human herpesvirus 6A and human herpesvirus 7 in AD brains (Readhead *et al.*, 2018). In addition, this thesis (chapter 3) identified genes involved with "interspecies interactions" were specifically enriched in AD brains when taking into account expression changes in related neurological disorders, findings which have been published in the Journal of Alzheimer's Disease (Patel, Dobson and Newhouse, 2019). The current observation extends the previous suggestions of viral involvement in AD brains to blood, as the expression of blood-derived genes involved in the "Herpes simplex infection" can be used to distinguish AD from other neurological diseases and control subjects in this study.

5.4.4 Predictive features consist of age-related markers

Age is one of the most significant risk factors for AD, and the prevalence of the disease is known to increase with age. A meta-analysis study investigating blood transcriptional changes associated with age in 14,983 humans, identified 1,496 differentially expressed genes with chronical age (Peters *et al.*, 2015), of which three genes (LDHB, AARS and ABR) are in the "AD vs mixed control" classification model's 28 predictive genes. The classification models most predictive gene LDHB is ranked 28th in the meta-analysis study and was also observed to be negatively associated with age in the brain, specifically the frontal cortex and cerebellum (Peters *et al.*, 2015). The datasets used in this study were publicly available, and as such, were accompanied with limited phenotypic information, including age. Therefore, age was not accounted for during the classification model developmental process. However, as this study uses a variety of age-related diseases, in addition to the 3 AD datasets, and study designs generally incorporate complementary age-matched controls, it is highly unlikely the classification model is predicting age alone but is more likely using a combination of signals including age to distinguish AD. Without age information for all subjects, this study is unable to conclude how age is influencing the model prediction process.

5.4.5 Gene expression profiling as a blood-derived biomarker for AD

A drive to identify accessible biomarkers for AD diagnosis has resulted in a field that has explored CSF and blood-derived components. Blood-derived biomarkers are particularly attractive due to their less invasive nature. Studies involving blood transcriptomic levels have demonstrated some capability of distinguishing AD from healthy controls (Fehlbaum-Beurdeley *et al.*, 2010; Booij *et al.*, 2011; Lunnon *et al.*, 2013; Roed *et al.*, 2013; Voyle *et al.*, 2016); however, small sample size and lack of independent validation datasets most likely led to overfitting and variable results. In addition replication of the underlying predictive genes across AD blood-based transcriptomic biomarker studies have also been found to be inconsistent (Siavelis *et al.*, 2016a), suggesting a robust, reproducible AD blood gene expression signature has still not been identified.

This study applied an alternative strategy to identify AD subjects by incorporating other diseases into the classification model developmental process, which in comparison to the typical AD vs healthy control approach, resulted in a model with an overall improved model performance when applied to a heterogeneous ageing population. Although the current clinical utility of the "AD vs mixed control" classification model is "excellent" in ruling out AD, the sensitivity of the test needs to be further improved for this type of research to be beneficial in the clinical setting. With the advances in new sequencing technologies and the ability to scan the whole transcriptome, additional gene-expression markers may increase the sensitivity performances whilst retaining specificity, providing a realistic

potential of blood-based transcriptomic classification model distinguishing AD from a heterogeneous ageing population.

5.4.6 Limitations

All data used in this study were publicly available, and as such, many were accompanied by limited phenotypic information, including basic sex information, which was predicted based on gene expression when missing. Therefore, this study was unable to incorporate additional phenotypic information during the classification model building process, which has been shown to improve model performance (Voyle *et al.*, 2016). Information such as comorbidities, age and medications are unknowns which could be affecting performance in this study. For instance, control subjects in this study that originated from non-AD datasets were screened negative for their corresponding disease of interest but were not screened for cognitive function. i.e. control subjects from the CD datasets were included in their retrospective dataset if they did not have CD, they were not necessarily checked for cognitive impairment. Therefore, some misclassified control subjects may indeed be on the AD spectrum, and it's important to note subjects from the pooled control group were most misclassified as AD by the "AD s mixed control" classification model. However, it is also important to note the training set used to develop the "AD vs mixed control" classification model also contains these controls which have not been screened for AD. If these controls or age-related disease subjects are comorbid with AD, the classification model may have inadvertently learned to be biased towards a subgroup of AD subjects with no comorbid with any other disease, hence the low sensitivity validation performance when introducing additional datasets into the classification model developmental process.

This study involved a number of subjects clinically diagnosed with age-related diseases, and most likely, are on some sort of therapeutic treatment to manage or treat the underlying disease, another piece of vital information generally missing from publicly available datasets and from this study. As therapeutic drugs have been well-known to affect gene expression profiling, including memantine, a common drug used to treat AD symptoms (Huang *et al.*, 2015), the "AD vs mixed control" classification model may have inadvertently learnt gene expression perturbations due to therapeutic treatment rather than disease biology, and would, therefore, fail in the clinical setting to diagnose AD subjects who are not already on medication. To address this issue along with co-morbidity, clear and detailed phenotypic information would be needed for all subjects, which is encouraged for future studies planning to submit genetic data to the public domain.

This study used datasets generated on 11 different microarray BeadArrays, resulting in datasets ranging from 22277-54715 probes prior to any QC. Coupled with differences in BeadArrays designs across platforms, the overlap of genes was drastically reduced to 1681 common "reliable detected" genes

across all datasets. This ensured gender-specific expression changes were captured; however, this may have also inadvertently lost some disease-specific changes. To address this issue, these subjects need to be expression profiled on the same microarray platform and ideally the same expression BeadArray, which currently doesn't exist in the public domain. The advances in sequencing technologies which can capture expression changes across the whole transcriptome can potentially solve this issue and future studies are encouraged to replicate this study design with RNA-Seq data with detailed phenotypic information when/if available, albeit, this may bring new challenges.

5.5 Conclusion

This study relied on publicly available microarray gene expression data, which too often lacks detailed phenotypic information for appropriate data analysis and needs to be addressed by future studies. Nevertheless, with the available phenotypic information and limited common "reliably detected" genes across the different microarray platforms and BeadArrays, this study, albeit it has many limitations, demonstrated the typical approach of developing an AD blood-based gene expression classification model using only AD and complimentary healthy controls fails to accurately distinguish AD from a heterogeneous ageing population. However, incorporating additional related neurological and age-related diseases into the classification model development process can result in a model with improved "predictive power" in distinguishing AD from a heterogeneous ageing population. Due to a number of limitations, further work is required in order to identify a robust blood transcriptomic signature more specific to AD.

Chapter 6: Discussion

6.1 Summary of principal findings

The aim of this thesis was to analyse inhouse and existing transcriptomic data to further understand the genetic and biological mechanisms associated with AD. The thesis began with chapter 2 investigating the first human brain microarray gene expression profiling of Asymptomatic AD (AsymAD) subjects who were clinically free from dementia; however, upon autopsy were discovered to be consistent with mild AD pathology. The aim of this chapter was to identify early brain transcriptomic changes and identify new potential therapeutic targets for early disease intervention. The third chapter combined novel and existing AD brain microarray gene expression studies in the largest known AD meta-analysis to date. Additional non-AD neurological disorders (PD, BD, Schizophrenia, MDD and HD) were incorporated into the analysis to identify specific molecular changes associated with AD brains. Furthermore, both brain regions spared and affected by hallmark AD pathology were analysed to reveal transcriptomic changes and disease mechanisms most likely involved with hallmark AD pathology.

The fourth chapter reprocessed the CMap database using up-to-date processing techniques and created an automated drug repositioning pipeline consisting of four different search algorithms. The AsymAD and AD brain transcriptomic signatures discovered in this thesis were repurposed to identify relevant candidate compounds that may intervene with the disease during the asymptomatic and symptomatic stage. Finally, in chapter 5, a machine learning approach was used in an attempt to identify a blood-based gene expression biomarker which could discriminate AD from cognitive healthy and age-related diseased subjects. The principal findings from each chapter are summarised in the following section.

6.1.1 Chapter 2: Transcriptomic Analysis of Probable Asymptomatic AD and AD Brains

A significant increase of transcriptomic activity in the FC brain region of AsymAD subjects was detected, suggesting fundamental changes in AD may initially begin within the FC brain region prior to clinical symptoms of AD. In addition, overactivation of the brain "glutamate-glutamine cycle" and disruption to the brain energy pathways in both AsymAD and AD subjects was detected. Further analysis using protein-protein interaction networks detected a shift from an already increased cell proliferation in AsymAD subjects to stress response and removal of amyloidogenic proteins in AD subjects.

6.1.2 Chapter 3: A Meta-analysis of Alzheimer's Disease Brain Transcriptomic Data

The meta-analysis identified 323, 435, 1023 and 828 DEG's specific to AD TL, FL, PL and CB brain regions respectively. Seven of these genes were consistently perturbed across all AD brain regions with SPCS1 gene expression pattern found to replicate in RNA-Seq data. The CB brain region was used as secondary

control to identify nineteen genes (ALDOA, GABBR1, TUBA1A, GAPDH, DNM3, KLC1, COX6C, ACTG1, CLTA, SLC25A5, PRNP, FDFT1, RHOQ, B2M, SPP1, WAC, UBA1, EIF4H and CLDND1) that may be involved in hallmark AD pathology. Furthermore, biological processes often reported as disrupted in AD were observed to be tissue-specific, with only the "metabolism of proteins" and viral components specifically enriched across AD brains.

6.1.3 Chapter 4: Automated Drug-repositioning from a disease expression signature

An automated drug repositioning pipeline was developed to query the reprocessed CMap database using a disease gene expression profile and four different algorithms to identify candidate compounds for disease intervention. Validation suggested three methods had the potential for drug repositioning; however, different methods were observed to be tissue-specific. Nevertheless, analysis of the CMap database identified tacrine (FDA approved for AD), thioperamide (FDA approved for cognitive impairment) and epirizole (FDA approved for anti-inflammatory disease) as candidate compounds for AsymAD but not AD patients. Furthermore, an anti-biotic candidate compound paromomycin was repurposed for AD. Through a literature review, the candidate compounds were found to be relevant to the disease; however, the work undertaken in this chapter was merely exploratory, and candidate compounds require further functional investigations.

6.1.4 Chapter 5: Working Towards a Blood-Derived Gene Expression Biomarker Specific for Alzheimer's Disease

The addition of related neurological and age-related disorders into the classification model development process identified a more AD-specific expression signature. The classification model achieved 66.3% PPV and 90.6% NPV when differentiating AD from Parkinson's Disease, Multiple Sclerosis, Amyotrophic Lateral Sclerosis, Bipolar Disorder, Schizophrenia, Coronary Artery Disease, Rheumatoid Arthritis, Chronic Obstructive Pulmonary Disease, and cognitively healthy subjects. Further investigation identified the predictive features were significantly enriched for "Herpes simplex infection" and age-related markers. Despite the improved specificity, a decrease in sensitivity was observed, and therefore, further improvement is still required in order to identify a reliable blood transcriptomic signature specific to AD.

6.2 Implications of findings

Much further work is still needed in this field to gain an insight into early and specific signatures of AD. Nevertheless, the findings from this thesis are discussed below.

6.2.1 Robust QC pipeline to process publicly available microarray gene expression data

This thesis provides a robust QC pipeline for processing raw Illumina and Affymetrix generated microarray gene expression data and makes it widely accessible as detailed in the "A Meta-Analysis of Alzheimer's Disease Brain Transcriptomic Data" peer-reviewed publication. Typically, an arbitrary cut-off or a fixed expression percentile threshold is used to determine if a gene is reliably detected, but this does not take into consideration the variation of probe quality across different platforms and expression arrays. As discovered in the "Chapter 5: Working Towards a Blood-Derived Gene Expression Biomarker Specific for Alzheimer's Disease", simply using common probes across different microarray platforms and expression arrays leads to variation in gene expression across samples, with obvious platform-specific batch effects observed. This can be largely explained by different platforms having different probe designs to target different transcripts of the same gene, leading to significant discrepancies and even absence in the measurement of the same gene by different platforms (Barnes *et al.*, 2005). However, when incorporating only probes deemed "reliably detected" by the data processing pipeline developed in this thesis, which uses a variety of literature defined house-keeping genes to determine if a gene is "reliably detected", this obvious platform-specific batch effect disappears. The application of this QC data processing pipeline is particularly useful when processing publicly available data, which as discovered in this thesis, is too often available in a "processed format only" where some probes have already been subset to those deemed "reliably detected" by corresponding authors.

6.2.2 Monitoring the FC brain region of the elderly may identify patients at risk of AD

Molecular changes in AD were found to initially begin in the FC brain region prior to the manifestation of clinical symptoms. This mirrors changes described in a longitudinal study involving ageing controls, where PET scans were used to detect increased activity in the medial frontal cortex in subjects who subsequently acquired cognitive impairment (Beason-Held *et al.*, 2013). Therefore, this thesis provides further evidence to clinically monitor brain activities in the FC of elderly patients to identify possible patients at risk of AD for early disease intervention.

6.2.3 Clinical treatment for AD may be useful in AsymAD

A significant overactivation of the "glutamate-glutamine cycle" in brains of probable asymptomatic AD subjects was identified, providing a possible therapeutic target for early disease intervention. Memantine is a compound that affects the "glutamate-glutamine cycle" and is already a clinically established treatment used for the symptomatic relief in AD patients. Memantine blocks N-methyl-d-aspartate (NMDA) receptors, preventing excitotoxicity caused by neurotransmitters such as glutamate, which temporarily increases cognition. In addition, repurposing the AsymAD expression signature in the

CMap database identified tacrine as one of the most significant candidate compounds, which once again is an FDA approved treatment for AD. Together, this suggests these compounds to treat AD may also be effective in the asymptomatic stage of the disease, although it's important to note identification of these patients in the symptomatic phase would be a challenge in itself, and these treatments do not change the pathology of the disease; however, they may possibly prolong the onset of disease symptoms and therefore require further investigation.

6.2.4 Identification of new genes associated with AD that may be of therapeutic benefit

Meta-analysis identified seven genes (down-regulated NDUFS5, SOD1, SPCS1 and up-regulated OGT, PURA, RERE, ZFP36L1) specifically perturbed across AD brains and not in similar brain regions of Schizophrenia, BD, HD, MDD or PD. The expression patterns of three of these genes (SPCS1, PURA and ZF36L1) were available in the RNA-Seq database Agora, which was found to relatively mirror microarray expression patterns. Subsequently, the AMP-AD consortia have nominated SPCS1 gene as a druggable target for AD based on expression patterns and additional network genomics and epigenomic elements. This provides confidence the remaining genes identified in this thesis may also be of therapeutic benefit and warrant further investigation.

6.2.5 Identification of new genes associated with hallmark AD pathology

Nineteen genes (ALDOA, GABBR1, TUBA1A, GAPDH, DNM3, KLC1, COX6C, ACTG1, CLTA, SLC25A5, PRNP, FDFT1, RHOQ, B2M, SPP1, WAC, UBA1, EIF4H and CLDND1) were identified to be associated specifically with hallmark AD pathology. Ten of these genes (DNM3, GABBR1, GAPDH, PRPN, FDFT1, KLC1, TUBA1A, CLTA, COX6C and SLC25A5) were previously reported to be associated with AD, with five genes (DNM3, GABBR1, GAPDH, PRPN and FDFT1) specifically suggested to be involved in the pathology of the disease. Therefore, this thesis further validates five genes are possibly associated with hallmark AD pathology, specifically NFT, whilst providing an additional fourteen candidate genes for further investigation.

6.2.6 Publicly available web-based application to explore gene expression changes in AD brains

The transcriptomic brain changes identified in this thesis which represent changes from healthy ageing to AsymAD and then further to AD, have been made publicly available through an easy to access R shiny web-based application (https://phidatalab-shiny.rosalind.kcl.ac.uk/ADbrainDE/). This allows future researchers to query specific genes and obtain expression patterns through the course of the disease and across various brain regions.

6.2.7 Gene expression data processing framework for the clinical environment

A microarray gene expression data processing framework was demonstrated in this thesis where datasets can be independently processed, and gene expression values directly compared to one another without the need to use additional information from a reference dataset to adjust gene expression values. This framework was demonstrated in the "Working Towards a Blood-Derived Gene Expression Biomarker Specific for Alzheimer's Disease" chapter where datasets in the training set were independently processed to build the classification model, and the validation set was independently processed and normalised without using gene expression variation information from the training set. This framework would allow for a real-world clinical application if such a transcriptomic based diagnostic test was developed, allowing new samples to be independently processed and diagnosis predicted without using prior expression variation knowledge from the data used to build the classification model, avoiding possible prediction biases.

6.2.8 Developing disease-specific classification models based on gene expression data

This thesis demonstrates to the research community the potential pitfalls of developing an AD blood-based gene expression classification model developed solely on cases and complimentary healthy controls. This typical approach may capture markers for illness and technical noise rather than biology specific to the disease, and as demonstrated in this thesis, when validating in an ageing heterogeneous population, the typical classification model fails to accurately distinguish AD from related disorders. However, incorporating related disorders into the classification model development process can aid in identifying a more AD-specific expression signature, with improved ability to distinguish AD from other related diseases and cognitively healthy controls. The approach used to build the classification model is most likely applicable to other disorders and can be used to refine a specific disease gene expression signatures.

6.2.9 A blood-based diagnostic test with excellent clinical utility for AD screening

The AD blood-based transcriptomic signature identified in this thesis has "excellent" clinical utility for screening; however, the utility is "poor" in identifying AD. Although improvement has been made when compared to the typical approach of building a blood-based AD classification model using simply AD subjects and complimentary healthy controls, further investigation is still required before a robust blood-based gene expression signature for clinical utility can be achieved, if any.

6.2.10 Bacterial and viral components in AD need further investigation

Throughout this thesis, bacterial and viral aspects have been identified to be associated with AD. Firstly, the meta-analysis identified various biological components enriched across all AD brain regions, however, when removing genes found to be perturbed in other neurological disorders, only four

biological components remain enriched and all four are indicative of interspecies interactions. Secondly, drug repositioning the AD brain gene expression signature identified paromomycin, an antibiotic to treat bacterial infections as the most significant candidate compound. Thirdly, the blood-based gene expression signature used to differentiate between AD from a heterogeneous ageing population was based on 28 predictive genes which were found to be most significantly enriched for "Herpes simplex infection". Aβ has been suggested to be an antimicrobial peptide and has been shown to protect against fungal and bacterial infections (Kumar *et al.*, 2016). Previous research has suggested bacterial involvement in AD (Sethi *et al.*, 2016; Readhead *et al.*, 2018), while another identified common viral species in normal and ageing brains, with an increased human herpesvirus 6A and human herpesvirus 7 in AD brains (Readhead *et al.*, 2018). The findings in this thesis reinforce bacterial and viral components may be involved in AD, prompting further investigation.

6.3 Limitations

The broader limitations of the thesis are presented here, with specific limitations addressed within individual chapters.

6.3.1 Microarray technologies

A big limitation of this thesis is the technology used to generate the data. Gene expression microarrays were designed to measure the expression of specific genes based on pre-defined probes and therefore are restricted to known variations. Therefore, microarrays may not be the ideal approach for transcriptomic biomarker discoveries. RNA-Seq provide many benefits over microarray technologies, including longer base pair reads for more precise mapping to the genome as microarray short reads can be limited due to the repetitiveness of the genome, has a larger dynamic range of expression measurement and has been suggested to be more reproducible (Wang, Gerstein and Snyder, 2009). However, as this thesis primarily relies on publicly available datasets, larger microarray datasets were more readily and abundantly available in comparison to RNA-Seq, making the analysis undertaken in this thesis possible.

6.3.2 Publicly available gene expression datasets lacked detailed phenotypic and sample processing information

Publicly available datasets lacked vital sample processing information, making it impossible to adjust for systematic errors introduced when samples are processed in multiple batches, a term often known as "batch effects". Therefore, this thesis incorporated best practices to estimate and correct for both known and hidden batch effects using SVA and COMBAT to ensure data is comparable between experiments and studies. However, this does not guarantee that all technical variation is completely removed.

In addition, many publicly available datasets lacked detailed phenotypic information such as gender, ethnicity, comorbidity, BRAAK staging, NFT, neuronal loss, age of disease onset, APOE status, rate of cognitive decline measures and disease duration. These influential factors, if known, could have been incorporated into the study design or data processing pipelines in an attempt to address their effects on the results.

Gene expression measurements can also be influenced by medication, which due to the lack of detailed phenotypic information accompanying datasets used in this thesis, it is unknown if subjects throughout this thesis were on any medication that may have influenced gene expression measurements. Clinically diagnosed AD subjects would have most likely been on medication for symptomatic relief through compounds such as memantine, which was found to influence gene expression in the "Automated Drug-repositioning of Transcriptomic Disease Signature" chapter. Therefore, the gene expression signatures captured in this thesis may be the result of the disease in combination with medication. Nevertheless, as none of the current FDA approved medications for AD affects the underlying pathology of the disease, the AD gene expression signatures identified in this thesis would have most likely captured biology relevant to AD, especially when associating with hallmark AD pathology.

6.4 Future directions

The "implications of findings" in combination with the thesis "limitations" sections outlined above have highlighted several future directions; however, a more general outline is provided below.

6.4.1 Validation of genes associated with AD

Several key genes perturbed in the brains of both asymptomatic and symptomatic AD have been identified using microarray gene expression datasets. Some of these genes have been previously associated with AD, while many are novel genes requiring further validation through additional technologies such as PCR or RNA-Seq. Providing these genes are found to be associated with AD, further investigations can attempt to identify the effects of these genes in AD through knockout mouse models.

6.4.2 An ensemble approach for drug repositioning and biomarker discovery

Drug repositioning was performed using four independent techniques to repurpose the CMap database. An ensemble of some of these methods may be more effective than the current single method approach and warrants further investigation. Similarly, the blood-based AD classification model was developed with a single machine learning approach; however, an ensemble of machine learning algorithms may be more effective in identifying a robust gene expression signature specific to AD as it has been demonstrated to improve predictive power (Zhou, 2012).

6.4.3 Blood-based gene expression disease multi-classifier

This thesis aimed to identify a blood-based gene expression signature to identify AD from an ageing heterogeneous population consisting of multiple disorders. Typically, a patient could be given such a blood test if symptoms were suggestive of dementia, which can be caused by many disorders including AD. A multi-classification model that can specifically identify or indicate the most likely cause of symptoms would be more beneficial. Therefore, future researchers are encouraged to explore this avenue using larger RNA-Seq datasets with detailed phenotypic information including comorbidity.

6.4.4 Biomarker for early detection of AD

The ability to detect AD early, during the asymptomatic phase of the disease would be extremely beneficial. This would allow an opportunity for the development of potential treatments to halt the disease early prior to clinical symptoms, which may be more effective than attempting to reverse the hallmark AD pathology that already exists in clinically diagnosed AD patients. Therefore, researchers are further encouraged to investigate biomarkers to identify these patients, which may be achieved through monitoring the neuronal activity of the FC brain region of elderly patients through imaging techniques such as PET.

6.4.5 Integration of biological, clinical and physical measurements to further advance the understanding of AD

The advances in technologies to measure an individual's biological and physical state provides an opportunity to combine genetics, transcriptomic, proteomic, brain imaging, physical state measurements from wearable devices and clinical symptoms to greatly advance the understanding of the risks, development and course of the disease. This can further help with the development of early diagnosis, define a definitive diagnosis of the disease or identify subgroups, and can provide an opportunity to provide individual orientated treatment.

6.4.6 Bacterial involvement in AD

With RNA-Seq technologies increasingly being used to profile AD and data being made publicly available, an opportunity may exist to investigate bacterial involvement in AD. One of the advantages of RNA-Seq over microarray technologies is the ability to be free from pre-defined probes, allowing measurement of annotation-independent transcription. Therefore, to infer biological relevance, the transcripts measured in RNA-Seq need to be aligned to a reference sequence. This provides an opportunity to align AD RNA-Seq transcripts to bacterial genomes to identify bacterial strains possibly involved in AD as suggest in (Croucher and Thomson, 2010).

6.5 Conclusion

The increase in life expectancy has profoundly increased the ageing population, which, unfortunately, is also accompanied by a rise in age-related disorders including AD. The most common form of dementia is AD, which was first described over a century ago, however, to date, there still exists a lack of understanding molecular changes specific to the disease, a clinically established robust blood-based biomarker for accurate disease diagnosis and a lack of therapeutic treatments.

This thesis was the first to explore the emergence of transcriptomic changes in the human brain from healthy ageing through to probable AsymAD and then further to AD. The results suggest molecular changes in AD may initially begin in the frontal cortex and disruptions in the energy cycle, and overactivation of the "glutamate-glutamine cycle" is already apparent in the asymptomatic phase of the disease. Compounds to target this pathway already exist. Memantine, an FDA approved treatment for AD, blocks NMDA receptors, preventing excitotoxicity caused by neurotransmitters such as glutamate, which can temporarily increase cognition. In addition, candidate compounds tacrine and thioperamide, which are also FDA approved to treat the symptoms of AD and cognitive impairment respectively, were repurposed for AsymAD in this thesis. This suggests FDA approved compounds to treat AD and general cognitive impairment may also be effective treatment options in the early phase of the disease, prior to clinical symptoms of AD, albeit the challenge remains to identify these patients and these compounds are known to only provide temporary symptomatic relief and do not change the underlying pathology of the disease.

The most extensive human AD microarray transcriptomic meta-analysis study to date was also performed in this thesis, which identified several genes specific to AD brains. Nineteen genes were discovered to be associated with hallmark AD pathology and seven genes were observed to be consistently perturbed across AD brains, with SPCS1 gene expression pattern found to replicate in RNA-Seq data. Subsequently, the AMP-AD consortia have nominated SPCS1 gene as a druggable target for AD based on expression patterns and additional network genomics and epigenomic elements. This provides additional confidence the remaining genes identified in this thesis may also be of therapeutic benefit.

Non-AD neurological disorders and age-related diseases were incorporated into this thesis to aid in the identification of biology that may be specific to AD, rather than general illness. This resulted in the discovery of bacterial and viral components being specifically enriched in the blood and brains of AD patients. Furthermore, repurposing the AD disease signature identified an antibiotic as the most significant candidate compound to treat AD. Together with recent literature, a bacterial and viral component may indeed exist in AD. In addition, this thesis demonstrated the addition of related

neurological and age-related diseases during the AD classification model development process results in a model with improved “predictive power” in distinguishing AD from a heterogeneous ageing population. However, further improvement is still required in order to identify a robust blood transcriptomic signature more specific to AD before clinical utility is considered.

Overall the work undertaken in this thesis provides new insight into the biological changes occurring in both the asymptomatic and symptomatic phase of the disease, demonstrates a framework for a possible blood-based transcriptomic diagnosis test, provides new potential therapeutic targets and identifies candidate compounds that require further investigation for AD intervention.

Bibliography

'2018 Alzheimer's disease facts and figures includes a special report on the financial and personal benefits of early diagnosis' (2018) *Alzheimers Dement*, 14(3), pp. 367–429. doi: 10.1016/j.jalz.2016.03.001.

Aarsland, D., Zaccai, J. and Brayne, C. (2005) 'A systematic review of prevalence studies of dementia in Parkinson's disease', *Movement Disorders*, 20(10), pp. 1255–1263. doi: 10.1002/mds.20527.

Aguet, F. *et al.* (2017) 'Genetic effects on gene expression across human tissues', *Nature*. Nature Publishing Group, 550(7675), pp. 204–213. doi: 10.1038/nature24277.

Albert, M. S. *et al.* (2011) 'The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease', *Alzheimer's & Dementia*, 7(3), pp. 270–279. doi: 10.1016/j.jalz.2011.03.008.

Allen, J. D., Chen, M. and Xie, Y. (2010) 'Model-Based Background Correction (MBCB): R Methods and GUI for Illumina Bead-array Data', *Journal of cancer science & therapy*. NIH Public Access, 1(1), pp. 25–27. doi: 10.4172/1948-5956.1000004.Model-Based.

Anand, R., Gill, K. D. and Mahdi, A. A. (2014) 'Therapeutics of Alzheimer's disease: Past, present and future', *Neuropharmacology*. Elsevier, 76, pp. 27–50. doi: 10.1016/j.neuropharm.2013.07.004.

Arevalo-Rodriguez, I. *et al.* (2015) 'Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI)', *Cochrane Database of Systematic Reviews*. John Wiley & Sons, Ltd, (3). doi: 10.1002/14651858.CD010783.pub2.

Armstrong, R. A. (2012) 'Visual signs and symptoms of dementia with Lewy bodies', *Clinical and Experimental Optometry*. John Wiley & Sons, Ltd (10.1111), 95(6), pp. 621–630. doi: 10.1111/j.1444-0938.2012.00770.x.

Bahr, T. M. *et al.* (2013) 'Peripheral blood mononuclear cell gene expression in chronic obstructive pulmonary disease', *American Journal of Respiratory Cell and Molecular Biology*, 49(2), pp. 316–323. doi: 10.1165/rcmb.2012-0230OC.

Balsis, S. *et al.* (2015) 'How Do Scores on the ADAS-Cog, MMSE, and CDR-SOB Correspond?', *The Clinical Neuropsychologist*, 29(7), pp. 1002–1009. doi: 10.1080/13854046.2015.1119312.

Barnes, M. *et al.* (2005) 'Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms', *Nucleic Acids Research*, 33(18), pp. 5914–5923. doi: 10.1093/nar/gki890.

Barrett, T. *et al.* (2013) 'NCBI GEO: archive for functional genomics data sets--update', *Nucleic Acids Research*, 41(D1), pp. D991–D995. doi: 10.1093/nar/gks1193.

Beason-Held, L. L. *et al.* (2013) 'Changes in Brain Function Occur Years before the Onset of Cognitive Impairment', *Journal of Neuroscience*, 33(46), pp. 18008–18014. doi: 10.1523/JNEUROSCI.1402-13.2013.

Beech, R. D. *et al.* (2010) 'Increased peripheral blood expression of electron transport chain genes in bipolar depression', *Bipolar Disorders*, 12(8), pp. 813–824. doi: 10.1111/j.1399-5618.2010.00882.x.

Behrens, C. J., van den Boom, L. P. and Heinemann, U. (2007) 'Effects of the $GABA_A$ receptor antagonists bicuculline and gabazine on stimulus-induced sharp wave-ripple complexes in adult rat

hippocampus *in vitro*', *European Journal of Neuroscience*. John Wiley & Sons, Ltd (10.1111), 25(7), pp. 2170–2181. doi: 10.1111/j.1460-9568.2007.05462.x.

Berchtold, N. C. *et al.* (2013) 'Synaptic genes are extensively downregulated across multiple brain regions in normal human aging and Alzheimer's disease', *Neurobiology of Aging*. Elsevier Inc, 34(6), pp. 1653–1661. doi: 10.1016/j.neurobiolaging.2012.11.024.

Besga, A. *et al.* (2015) 'Discrimination between Alzheimer's Disease and Late Onset Bipolar Disorder Using Multivariate Analysis.', *Frontiers in aging neuroscience*. Frontiers Media SA, 7, p. 231. doi: 10.3389/fnagi.2015.00231.

van Beveren, N. J. M. *et al.* (2012) 'Marked reduction of AKT1 expression and deregulation of AKT1-associated pathways in peripheral blood mononuclear cells of schizophrenia patients', *PLoS ONE*, 7(2). doi: 10.1371/journal.pone.0032618.

Bipolar disorder - NHS (no date). Available at: https://www.nhs.uk/conditions/bipolar-disorder/ (Accessed: 31 January 2019).

Bird, T. D. (2018) *Alzheimer Disease Overview*. Available at: https://www.ncbi.nlm.nih.gov/books/NBK1161/ (Accessed: 2 January 2019).

Birks, J. S. (2006) 'Cholinesterase inhibitors for Alzheimer's disease', in Birks, J. S. (ed.) *Cochrane Database of Systematic Reviews*. Chichester, UK: John Wiley & Sons, Ltd, p. CD005593. doi: 10.1002/14651858.CD005593.

Bittner, T. *et al.* (2016) 'Technical performance of a novel, fully automated electrochemiluminescence immunoassay for the quantitation of β-amyloid (1-42) in human cerebrospinal fluid'. doi: 10.1016/j.jalz.2015.09.009.

Blalock, E. M. *et al.* (2004) 'Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses', *Proceedings of the National Academy of Sciences*, 101(7), pp. 2173–2178. doi: 10.1073/pnas.0308512100.

Blalock, E. M. *et al.* (2011) 'Microarray analyses of laser-captured hippocampus reveal distinct gray and white matter signatures associated with incipient Alzheimer's disease', *Journal of Chemical Neuroanatomy*. Elsevier B.V., 42(2), pp. 118–126. doi: 10.1016/j.jchemneu.2011.06.007.

Blennow, K. (2017) 'A Review of Fluid Biomarkers for Alzheimer's Disease: Moving from CSF to Blood.', *Neurology and therapy*. Springer, 6(Suppl 1), pp. 15–24. doi: 10.1007/s40120-017-0073-9.

Bolós, M., Perea, J. R. and Avila, J. (2017) 'Alzheimer's disease as an inflammatory disease', *Biomolecular Concepts*. De Gruyter, 8(1), pp. 37–43. doi: 10.1515/bmc-2016-0029.

Booij, B. B. *et al.* (2011) 'A gene expression pattern in blood for the early detection of Alzheimer's disease', *Journal of Alzheimer's Disease*, 23(1), pp. 109–119. doi: 10.3233/JAD-2010-101518.

Braak, H. and Braak, E. (1991) 'Neuropathological stageing of Alzheimer-related changes', *Acta Neuropathologica*. Springer-Verlag, 82(4), pp. 239–259. doi: 10.1007/BF00308809.

Brazma, A. *et al.* (2003) 'ArrayExpress - A public repository for microarray gene expression data at the EBI', *Nucleic Acids Research*, 31(1), pp. 68–71. doi: 10.1093/nar/gkg091.

Breitner, J. C. *et al.* (2011) 'Extended results of the Alzheimer's disease anti-inflammatory prevention trial.', *Alzheimer's & dementia : the journal of the Alzheimer's Association*. NIH Public Access, 7(4), pp. 402–11. doi: 10.1016/j.jalz.2010.12.014.

Buckberry, S. *et al.* (2014) 'MassiR: A method for predicting the sex of samples in gene expression microarray datasets', *Bioinformatics*, 30(14), pp. 2084–2085. doi: 10.1093/bioinformatics/btu161.

Calciano, M. a *et al.* (2010) 'Drug treatment of Alzheimer's disease patients leads to expression changes in peripheral blood cells.', *Alzheimer's & dementia : the journal of the Alzheimer's Association*. Elsevier Ltd, 6(5), pp. 386–393. doi: 10.1016/j.jalz.2009.12.004.

Calligaris, R. *et al.* (2015) 'Blood transcriptomics of drug-naïve sporadic Parkinson's disease patients', *BMC Genomics*. BMC Genomics, 16(1), pp. 1–14. doi: 10.1186/s12864-015-2058-3.

Cardenas, V. A. *et al.* (2011) 'Brain atrophy associated with baseline and longitudinal measures of cognition', *Neurobiology of Aging*. Elsevier, 32(4), pp. 572–580. doi: 10.1016/J.NEUROBIOLAGING.2009.04.011.

Caselli, R. J. and Reiman, E. M. (2012) 'Characterizing the preclinical stages of Alzheimer's disease and the prospect of presymptomatic intervention', *Advances in Alzheimer's Disease*, 3(0 1), pp. 405–416. doi: 10.3233/978-1-61499-154-0-405.

Cassimeris, L. (2009) 'Microtubule Associated Proteins in Neurons', *Encyclopedia of Neuroscience*. Academic Press, pp. 865–870. doi: 10.1016/B978-008045046-9.00725-7.

Chandrasekaran, S. and Bonchev, D. (2013) 'A network view on Parkinson's disease.', *Computational and structural biotechnology journal*, 7(8), p. e201304004. doi: 10.5936/csbj.201304004.

Chang, C. W. *et al.* (2011) 'Identification of human housekeeping genes and Tissue-Selective genes by microarray Meta-Analysis', *PLoS ONE*, 6(7), pp. 14–19. doi: 10.1371/journal.pone.0022859.

Chang, L.-C. *et al.* (2013) 'Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline', *BMC Bioinformatics*. BioMed Central, 14(1), p. 368. doi: 10.1186/1471-2105-14-368.

Chen, C. *et al.* (2011) 'Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods.', *PloS one*. Edited by D. Kliebenstein. Public Library of Science, 6(2), p. e17238. doi: 10.1371/journal.pone.0017238.

Chen, F. *et al.* (2013) 'Gene expression profile and functional analysis of Alzheimer's disease.', *American journal of Alzheimer's disease and other dementias*, 28(7), pp. 693–701. doi: 10.1177/1533317513500838.

Chen, H. *et al.* (2013) 'Gene expression alterations in bipolar disorder postmortem brains', *Bipolar Disorders*, 15(2), pp. 177–187. doi: 10.1111/bdi.12039.

Chen, J. *et al.* (2016) 'Gene expression analysis reveals the dysregulation of immune and metabolic pathways in Alzheimer’s disease', *Oncotarget*, 7(45), pp. 1–6. doi: 10.18632/oncotarget.12505.

Chen, J. J. *et al.* (2007) 'Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data', *BMC Bioinformatics*. BioMed Central, 8(1), p. 412. doi: 10.1186/1471-2105-8-412.

Chen, L. *et al.* (2017) 'Correlation between RNA-Seq and microarrays results using TCGA data', *Gene*. Elsevier, 628, pp. 200–204. doi: 10.1016/J.GENE.2017.07.056.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A Scalable Tree Boosting System'. doi: 10.1145/2939672.2939785.

Chiam, J. T. W. *et al.* (2014) 'Are Blood-Based Protein Biomarkers for Alzheimer's Disease also Involved in Other Brain Disorders? A Systematic Review.', *Journal of Alzheimer's disease : JAD*, 43, pp. 303–314. doi: 10.3233/JAD-140816.

De Chiara, G. *et al.* (2012) 'Infectious agents and neurodegeneration', *Molecular Neurobiology*,

46(3), pp. 614–638. doi: 10.1007/s12035-012-8320-7.

Clelland, C. L. *et al.* (2013) 'Utilization of Never-Medicated Bipolar Disorder Patients towards Development and Validation of a Peripheral Biomarker Profile', *PLoS ONE*, 8(6), pp. 1–11. doi: 10.1371/journal.pone.0069082.

Coan, G. and Mitchell, C. S. (2015) 'An Assessment of Possible Neuropathology and Clinical Relationships in 46 Sporadic Amyotrophic Lateral Sclerosis Patient Autopsies', *Neurodegenerative Diseases*. Karger Publishers, 15(5), pp. 301–312. doi: 10.1159/000433581.

Cohen, H. *et al.* (2006) 'Anisomycin, a Protein Synthesis Inhibitor, Disrupts Traumatic Memory Consolidation and Attenuates Posttraumatic Stress Response in Rats', *Biological Psychiatry*, 60(7), pp. 767–776. doi: 10.1016/j.biopsych.2006.03.013.

Colangelo, V. *et al.* (2002) 'Gene expression profiling of 12633 genes in Alzheimer hippocampal CA1: Transcription and neurotrophic factor down-regulation and up-regulation of apoptotic and pro-inflammatory signaling', *Journal of Neuroscience Research*, 70(3), pp. 462–473. doi: 10.1002/jnr.10351.

Consortium, M. (2008) 'The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements', 42(2), pp. 157–162. doi: 10.1037/a0030561.Striving.

Convit, a. *et al.* (2000) 'Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease☆', *Neurobiology of Aging*, 21(1), pp. 19–26. doi: 10.1016/S0197-4580(99)00107-4.

Corbett, A. *et al.* (2012) 'Drug repositioning for Alzheimer's disease', *Nature Reviews Drug Discovery*. Nature Publishing Group, 11(11), pp. 833–846. doi: 10.1038/nrd3869.

Crismon, M. L. (1994) 'Tacrine: First Drug Approved for Alzheimer's Disease', *Annals of Pharmacotherapy*. SAGE Publications, 28(6), pp. 744–751. doi: 10.1177/106002809402800612.

Croucher, N. J. and Thomson, N. R. (2010) 'Studying bacterial transcriptomes using RNA-seq', *Current Opinion in Microbiology*. Elsevier Current Trends, 13(5), pp. 619–624. doi: 10.1016/J.MIB.2010.09.009.

Cummings, J., Feldman, H. H. and Scheltens, P. (2019) 'The "rights" of precision drug development for Alzheimer's disease', *Alzheimer's Research & Therapy*. Springer Science and Business Media LLC, 11(1). doi: 10.1186/s13195-019-0529-5.

Cummings, J., Reiber, C. and Kumar, P. (2018) 'The price of progress: Funding and financing Alzheimer's disease drug development.', *Alzheimer's & dementia (New York, N. Y.)*. Elsevier, 4, pp. 330–343. doi: 10.1016/j.trci.2018.04.008.

Cunnane, S. *et al.* (2011) 'Brain fuel metabolism, aging, and Alzheimer's disease', *Nutrition*, 27(1), pp. 3–20. doi: 10.1016/j.nut.2010.07.021.

Dal Bianco, A. *et al.* (2008) 'Multiple sclerosis and Alzheimer's disease', *Annals of Neurology*, 63(2), pp. 174–183. doi: 10.1002/ana.21240.

Davis, M. A. *et al.* (2015) 'HHS Public Access', 175(5), pp. 777–783. doi: 10.1001/jamainternmed.2014.5466.Association.

Dekker, A. D. *et al.* (2017) 'Cerebrospinal fluid biomarkers for Alzheimer's disease in Down syndrome.', *Alzheimer's & dementia (Amsterdam, Netherlands)*. Elsevier, 8, pp. 1–10. doi: 10.1016/j.dadm.2017.02.006.

DeMichele-Sweet, M. A. A. *et al.* (2018) 'Genetic risk for schizophrenia and psychosis in Alzheimer disease.', *Molecular psychiatry*. NIH Public Access, 23(4), pp. 963–972. doi: 10.1038/mp.2017.81.

Dhaliwal, S. *et al.* (2018) 'Effective Intrusion Detection System Using XGBoost', *Information*. Multidisciplinary Digital Publishing Institute, 9(7), p. 149. doi: 10.3390/info9070149.

Dileep, K. V. *et al.* (2013) 'Designing of multi-target-directed ligands against the enzymes associated with neuroinflammation: an *in silico* approach', *Frontiers in Life Science*. Taylor & Francis, 7(3–4), pp. 174–185. doi: 10.1080/21553769.2014.901924.

Ding, L.-H. *et al.* (2008) 'Enhanced identification and biological validation of differential gene expression via Illumina whole-genome expression arrays through the use of the model-based background correction methodology', *Nucleic Acids Research*. Oxford University Press, 36(10), pp. e58–e58. doi: 10.1093/nar/gkn234.

Dominy, S. S. *et al.* (2019) '*Porphyromonas gingivalis* in Alzheimer's disease brains: Evidence for disease causation and treatment with small-molecule inhibitors', *Science Advances*. American Association for the Advancement of Science, 5(1), p. eaau3333. doi: 10.1126/sciadv.aau3333.

Driscoll, I. and Troncoso, J. (2011) 'Asymptomatic Alzheimer's disease: a prodrome or a state of resilience?', *Current Alzheimer research*, 8(4), pp. 330–5. doi: 10.2174/1567211212225942050.

Du, P., Kibbe, W. a and Lin, S. M. (2008) 'lumi: a pipeline for processing Illumina microarray.', *Bioinformatics (Oxford, England)*, 24(13), pp. 1547–8. doi: 10.1093/bioinformatics/btn224.

Duan, Q. *et al.* (2014) 'LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures.', *Nucleic acids research*. Oxford University Press, 42(Web Server issue), pp. W449-60. doi: 10.1093/nar/gku476.

Dubois, B. *et al.* (2015) 'Timely Diagnosis for Alzheimer's Disease: A Literature Review on Benefits and Challenges', *Journal of Alzheimer's Disease*. Edited by A. Saykin, 49(3), pp. 617–631. doi: 10.3233/JAD-150692.

Edwards, Y. J. K. *et al.* (2011) 'Identifying consensus disease pathways in Parkinson's disease using an integrative systems biology approach', *PLoS ONE*, 6(2). doi: 10.1371/journal.pone.0016917.

Eimer, W. A. *et al.* (2018) 'Alzheimer's Disease-Associated β-Amyloid Is Rapidly Seeded by Herpesviridae to Protect against Brain Infection.', *Neuron*. NIH Public Access, 99(1), pp. 56-63.e3. doi: 10.1016/j.neuron.2018.06.030.

Ein-Dor, L. *et al.* (2005) 'Outcome signature genes in breast cancer: is there a unique set?', *Bioinformatics*, 21(2), pp. 171–178. doi: 10.1093/bioinformatics/bth469.

Ellsworth, D. L. *et al.* (2014) 'Intensive cardiovascular risk reduction induces sustainable changes in expression of genes and pathways important to vascular function', *Circulation: Cardiovascular Genetics*, 7(2), pp. 151–160. doi: 10.1161/CIRCGENETICS.113.000121.

Escott-Price, V. *et al.* (2015) 'Common polygenic variation enhances risk prediction for Alzheimer's disease', *Brain*, 138(12), pp. 3673–3684. doi: 10.1093/brain/awv268.

Fayed, N. *et al.* (2011) 'Brain glutamate levels are decreased in alzheimer's disease: A magnetic resonance spectroscopy study', *American Journal of Alzheimer's Disease and other Dementias*, 26(6), pp. 450–456. doi: 10.1177/1533317511421780.

Fehlbaum-Beurdeley, P. *et al.* (2010) 'Toward an Alzheimer's disease diagnosis via high-resolution blood gene expression', *Alzheimer's and Dementia*, 6, pp. 25–38. doi: 10.1016/j.jalz.2009.07.001.

Fleisher, A. S. *et al.* (2011) 'Using Positron Emission Tomography and Florbetapir F 18 to Image

Cortical Amyloid in Patients With Mild Cognitive Impairment or Dementia Due to Alzheimer Disease', *Archives of Neurology*, 68(11), p. 1404. doi: 10.1001/archneurol.2011.150.

Foiani, M. S. *et al.* (2018) 'Plasma tau is increased in frontotemporal dementia.', *Journal of neurology, neurosurgery, and psychiatry*. BMJ Publishing Group Ltd, 89(8), pp. 804–807. doi: 10.1136/jnnp-2017-317260.

Folstein, M. F., Folstein, S. E. and McHugh, P. R. (1975) '"Mini-mental state": A practical method for grading the cognitive state of patients for the clinician', *Journal of Psychiatric Research*. Pergamon, 12(3), pp. 189–198. doi: 10.1016/0022-3956(75)90026-6.

Forlenza, O. V. and Aprahamian, I. (2013) 'Cognitive impairment and dementia in bipolar disorder.', *Frontiers in bioscience (Elite edition)*, 5, pp. 258–65. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23276987 (Accessed: 31 January 2019).

Fortney, K. *et al.* (2015) 'Prioritizing Therapeutics for Lung Cancer: An Integrative Meta-analysis of Cancer Gene Signatures and Chemogenomic Data', *PLoS Computational Biology*. Public Library of Science, 11(3), pp. 1–17. doi: 10.1371/journal.pcbi.1004068.

Fowler, K. D. *et al.* (2015) 'Leveraging existing data sets to generate new insights into Alzheimer's disease biology in specific patient subsets', *Scientific Reports*. Nature Publishing Group, 5(1), p. 14324. doi: 10.1038/srep14324.

Gaugler, J. *et al.* (2016) '2016 Alzheimer's disease facts and figures', *Alzheimer's and Dementia*, 12(4), pp. 459–509. doi: 10.1016/j.jalz.2016.03.001.

Godoy, J. A. *et al.* (2014) 'Signaling pathway cross talk in Alzheimer's disease', *Cell Communication and Signaling*, 12(1), p. 23. doi: 10.1186/1478-811X-12-23.

Goertsches, R. H. *et al.* (2010) 'Long-term genome-wide blood RNA expression profiles yield novel molecular response candidates for IFN-??-1b treatment in relapsing remitting MS', *Pharmacogenomics*, 11(2), pp. 147–161. doi: 10.2217/pgs.09.152.

Gogtay, N. *et al.* (2011) 'Age of onset of schizophrenia: perspectives from structural neuroimaging studies.', *Schizophrenia bulletin*. Oxford University Press, 37(3), pp. 504–13. doi: 10.1093/schbul/sbr030.

Goldenberg, M. M. (2012) 'Multiple sclerosis review.', *P & T : a peer-reviewed journal for formulary management*. MediMedia, USA, 37(3), pp. 175–84. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22605909 (Accessed: 3 February 2019).

Grove, R. *et al.* (2014) 'A Randomized, Double-Blind, Placebo-Controlled, 16-Week Study of the H_3 Receptor Antagonist, GSK239512 as a Monotherapy in Subjects with Mild-to-Moderate Alzheimer's Disease', *Current Alzheimer Research*, 11(1), pp. 47–58. doi: 10.2174/1567205010666131212110148.

Grzywacz, J. G., Segel-Karpas, D. and Lachman, M. E. (2016) 'Workplace Exposures and Cognitive Function During Adulthood: Evidence From National Survey of Midlife Development and the O*NET.', *Journal of occupational and environmental medicine*. NIH Public Access, 58(6), pp. 535–41. doi: 10.1097/JOM.0000000000000727.

Guiroy, D. C. *et al.* (1993) 'Neurofibrillary tangles of Guamanian amyotrophic lateral sclerosis, parkinsonism-dementia and neurologically normal Guamanians contain a 4- to 4.5-kilodalton protein which is immunoreactive to anti-amyloid beta/A4-protein antibodies.', *Acta neuropathologica*, 86(3), pp. 265–74. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8213085 (Accessed: 31 January 2019).

Guskiewicz, K. M. *et al.* (2005) 'Association between recurrent concussion and late-life cognitive impairment in retired professional football players.', *Neurosurgery*, 57(4), pp. 719–26; discussion 719-26. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16239884 (Accessed: 2 January 2019).

Guyon, I. *et al.* (2013) 'Tracking cellulase behaviors', *Biotechnology and Bioengineering*, 110(1), p. fmvi-fmvi. doi: 10.1002/bit.24634.

Gyorffy, B. *et al.* (2009) 'Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples.', *PloS one*. Public Library of Science, 4(5), p. e5645. doi: 10.1371/journal.pone.0005645.

Hackstadt, A. J. and Hess, A. M. (2009) 'Filtering for increased power for microarray data analysis', *BMC Bioinformatics*, 10(1), p. 11. doi: 10.1186/1471-2105-10-11.

Hampel, H. *et al.* (2018) 'Blood-based biomarkers for Alzheimer disease: mapping the road to the clinic', *Nature Reviews Neurology*. Nature Publishing Group, 14(11), pp. 639–652. doi: 10.1038/s41582-018-0079-7.

Han, G. *et al.* (2013) 'Characteristic transformation of blood transcriptome in Alzheimer's disease', *Journal of Alzheimer's Disease*, 35(2), pp. 373–86. doi: 10.3233/JAD-121963.

Harr, B. and Schlötterer, C. (2006) 'Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons.', *Nucleic acids research*. Oxford University Press, 34(2), p. e8. doi: 10.1093/nar/gnj010.

Harrison, P. J., Colbourne, L. and Harrison, C. H. (2018) 'The neuropathology of bipolar disorder: systematic review and meta-analysis', *Molecular Psychiatry*. Nature Publishing Group, p. 1. doi: 10.1038/s41380-018-0213-3.

Haury, A.-C., Gestraud, P. and Vert, J.-P. (2011) 'The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures', *PLoS ONE*. Edited by M.-T. Teh. Public Library of Science, 6(12), p. e28210. doi: 10.1371/journal.pone.0028210.

He, R. *et al.* (2015) 'Cefsulodin Inspired Potent and Selective Inhibitors of mPTPB, a Virulent Phosphatase from Mycobacterium tuberculosis.', *ACS medicinal chemistry letters*. American Chemical Society, 6(12), pp. 1231–5. doi: 10.1021/acsmedchemlett.5b00373.

Henriksen, K. *et al.* (2014) 'The future of blood-based biomarkers for Alzheimer's disease', *Alzheimer's and Dementia*, 10, pp. 115–131. doi: 10.1016/j.jalz.2013.01.013.

Hilal, S. *et al.* (2017) 'Plasma Amyloid-β Levels, Cerebral Small Vessel Disease, and Cognition: The Rotterdam Study', *Journal of Alzheimer's Disease*. Edited by L. Zahodne, 60(3), pp. 977–987. doi: 10.3233/JAD-170458.

Hippius, H. and Neundörfer, G. (2003) 'The discovery of Alzheimer's disease', *Dialogues in Clinical Neuroscience*, 5(1), pp. 101–108. doi: 10.1055/s-2008-1026558.

Hokama, M. *et al.* (2014) 'Altered expression of diabetes-related genes in Alzheimer's disease brains: the Hisayama study.', *Cerebral cortex (New York, N.Y. : 1991)*, 24(9), pp. 2476–88. doi: 10.1093/cercor/bht101.

Hong, L., Huang, H.-C. and Jiang, Z.-F. (2014) 'Relationship between amyloid-beta and the ubiquitin–proteasome system in Alzheimer's disease', *Neurological Research*, 36(3), pp. 276–282. doi: 10.1179/1743132813Y.0000000288.

Howard, R. (2000) 'Late-Onset Schizophrenia and Very-Late-Onset Schizophrenia-Like Psychosis: An International Consensus', *American Journal of Psychiatry*, 157(2), pp. 172–178. doi: 10.1176/appi.ajp.157.2.172.

Huang, H. *et al.* (2015) 'DMAP : a connectivity map database to enable identification of novel drug repositioning candidates', *BMC Bioinformatics*. BioMed Central Ltd, 16(Suppl 13), p. S4. doi: 10.1186/1471-2105-16-S13-S4.

Hunt, H. A. *et al.* (2017) 'The comparative diagnostic accuracy of the Mini Mental State Examination (MMSE) and the General Practitioner assessment of Cognition (GPCOG) for identifying dementia in primary care: a systematic review protocol', *Diagnostic and Prognostic Research*. BioMed Central, 1(1), p. 14. doi: 10.1186/s41512-017-0014-1.

Huynh, R. A. and Mohan, C. (2017) 'Alzheimer's disease: Biomarkers in the genome, blood, and cerebrospinal fluid', *Frontiers in Neurology*, 8(MAR). doi: 10.3389/fneur.2017.00102.

Iacono, D. *et al.* (2015) 'APOε2 and education in cognitively normal older subjects with high levels of AD pathology at autopsy : findings from the Nun Study', *Oncotarget*, 6(16). doi: https://doi.org/10.18632/oncotarget.4118.

Ikonomovic, M. D. *et al.* (2008) 'Post-mortem correlates of in vivo PiB-PET amyloid imaging in a typical case of Alzheimer's disease', *Brain*, 131(6), pp. 1630–1645. doi: 10.1093/brain/awn016.

Inoue, K. *et al.* (2013) 'Metabolic profiling of Alzheimer's disease brains', *Scientific Reports*. Nature Publishing Group, 3(1), p. 2364. doi: 10.1038/srep02364.

Iorio, F. *et al.* (2013) 'Transcriptional data: a new gateway to drug repositioning?', *Drug Discovery Today*. Elsevier Ltd, 18(7–8), pp. 350–7. doi: 10.1016/j.drudis.2012.07.014.

Iqbal, K. *et al.* (2010) 'Tau in Alzheimer disease and related tauopathies.', *Current Alzheimer research*. NIH Public Access, 7(8), pp. 656–64. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20678074 (Accessed: 14 January 2019).

Iritani, S. (2007) 'Neuropathology of schizophrenia: A mini review', *Neuropathology*, 27(6), pp. 604–608. doi: 10.1111/j.1440-1789.2007.00798.x.

Irizar, H. *et al.* (2014) 'Transcriptomic profile reveals gender-specific molecular mechanisms driving multiple sclerosis progression', *PLoS ONE*, 9(2). doi: 10.1371/journal.pone.0090482.

Irizarry, R. A. *et al.* (2003) 'Exploration, normalization, and summaries of high density oligonucleotide array probe level data', *Biostatistics*. Oxford University Press, 4(2), pp. 249–264. doi: 10.1093/biostatistics/4.2.249.

Ishii, K. *et al.* (1997) 'Reduction of cerebellar glucose metabolism in advanced Alzheimer's disease.', *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, 38(6), pp. 925–928. Available at: http://jnm.snmjournals.org/content/38/6/925.full.pdf.

Itzhaki, R. F. *et al.* (1997) 'Herpes simplex virus type 1 in brain and risk of Alzheimer's disease', *The Lancet*, 349(9047), pp. 241–244. doi: 10.1016/S0140-6736(96)10149-5.

Itzhaki, R. F. (2018) 'Corroboration of a Major Role for Herpes Simplex Virus Type 1 in Alzheimer's Disease', *Frontiers in Aging Neuroscience*. Frontiers, 10, p. 324. doi: 10.3389/fnagi.2018.00324.

Jacobs, H. I. L. *et al.* (2017) 'The cerebellum in Alzheimer's disease: evaluating its role in cognitive decline', *Brain*, (January). doi: 10.1093/brain/awx194.

De Jager, P. L. *et al.* (2009) 'Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci', *Nature Genetics*. Nature Publishing Group, 41(7), pp. 776–782. doi: 10.1038/ng.401.

Jiang, N. *et al.* (2008) 'Methods for evaluating gene expression from Affymetrix microarray datasets.', *BMC bioinformatics*. BioMed Central, 9, p. 284. doi: 10.1186/1471-2105-9-284.

Johnson, J. W. and Kotermanski, S. E. (2006) 'Mechanism of action of memantine', *Current Opinion in Pharmacology*, 6(1), pp. 61–67. doi: 10.1016.

de Jong, S. *et al.* (2012) 'A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes.', *PloS one*, 7(6), p. e39498. doi: 10.1371/journal.pone.0039498.

El Kadmiri, N. *et al.* (2018) 'Biomarkers for Alzheimer Disease: Classical and Novel Candidates' Review', *Neuroscience*. Pergamon, 370, pp. 181–190. doi: 10.1016/J.NEUROSCIENCE.2017.07.017.

Kamburov, A. *et al.* (2009) 'ConsensusPathDB - A database for integrating human functional interaction networks', *Nucleic Acids Research*, 37(SUPPL. 1), pp. 623–628. doi: 10.1093/nar/gkn698.

Kametani, F. and Hasegawa, M. (2018) 'Reconsideration of Amyloid Hypothesis and Tau Hypothesis in Alzheimer's Disease.', *Frontiers in neuroscience*. Frontiers Media SA, 12, p. 25. doi: 10.3389/fnins.2018.00025.

Kang, D. D. *et al.* (2012) 'MetaQC: Objective quality control and inclusion/exclusion criteria for genomic meta-analysis', *Nucleic Acids Research*, 40(2), pp. 1–14. doi: 10.1093/nar/gkr1071.

Khachaturian, A. S. *et al.* (2018) 'Future prospects and challenges for Alzheimer's disease drug development in the era of the NIA-AA Research Framework', *Alzheimer's and Dementia*. Elsevier Inc., pp. 532–534. doi: 10.1016/j.jalz.2018.03.003.

Khanzada, N., Butler, M. and Manzardo, A. (2017) 'GeneAnalytics Pathway Analysis and Genetic Overlap among Autism Spectrum Disorder, Bipolar Disorder and Schizophrenia', *International Journal of Molecular Sciences*, 18(3), p. 527. doi: 10.3390/ijms18030527.

Kiddle, S. J. *et al.* (2014) 'Candidate blood proteome markers of Alzheimer's disease onset and progression: a systematic review and replication study.', *Journal of Alzheimer's disease : JAD*, 38(3), pp. 515–31. doi: 10.3233/JAD-130380.

Kiyohara, C. and Yoshimasu, K. (2009) 'Molecular epidemiology of major depressive disorder.', *Environmental health and preventive medicine*. BioMed Central, 14(2), pp. 71–87. doi: 10.1007/s12199-008-0073-6.

Kokubo, Y. and Kuzuhara, S. (2004) 'Neurofibrillary tangles in ALS and Parkinsonism-dementia complex focus in Kii, Japan.', *Neurology*. Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology, 63(12), pp. 2399–401. doi: 10.1212/01.WNL.0000147241.52694.6A.

Köpke, E. *et al.* (1993) 'Microtubule-associated protein tau. Abnormal phosphorylation of a non-paired helical filament pool in Alzheimer disease.', *The Journal of biological chemistry*, 268(32), pp. 24374–84. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8226987 (Accessed: 14 January 2019).

Kumar, D. K. V. *et al.* (2016) 'Amyloid- peptide protects against microbial infection in mouse and worm models of Alzheimers disease', *Science Translational Medicine*, 8(340), pp. 340ra72-340ra72. doi: 10.1126/scitranslmed.aaf1059.

Kwakowsky, A. *et al.* (2018) 'Gamma-aminobutyric acid A receptors in Alzheimer's disease: highly localized remodeling of a complex and diverse signaling pathway.', *Neural regeneration research*. Wolters Kluwer -- Medknow Publications, 13(8), pp. 1362–1363. doi: 10.4103/1673-5374.235240.

Labadorf, A. *et al.* (2015) 'RNA sequence analysis of human huntington disease brain reveals an extensive increase in inflammatory and developmental gene expression', *PLoS ONE*, 10(12), pp. 1–21. doi: 10.1371/journal.pone.0143563.

Lamb, J. *et al.* (2006) 'The Connectivity Map : Using gene-expression signatures to connect small

molecules, genes, and disease.', *Science*, 313(September), pp. 1929–1935. doi: 10.1126/science.1132939.

Lambert, J. C. *et al.* (2010) 'Implication of the immune system in Alzheimer's disease: evidence from genome-wide pathway analysis', *Journal of Alzheimer's Disease*, 20(4), pp. 1107–1118. doi: 10.3233/JAD-2010-100018.

Langfelder, P. *et al.* (2011) 'Is my network module preserved and reproducible?', *PLoS Computational Biology*, 7(1). doi: 10.1371/journal.pcbi.1001057.

Langfelder, P. and Horvath, S. (2008) 'WGCNA: an R package for weighted correlation network analysis.', *BMC bioinformatics*, 9, p. 559. doi: 10.1186/1471-2105-9-559.

Lazar, C. *et al.* (2013) 'Batch effect removal methods for microarray gene expression data integration: A survey', *Briefings in Bioinformatics*, 14, pp. 469–490. doi: 10.1093/bib/bbs037.

Lê Cao, K.-A. *et al.* (2014) 'YuGene: a simple approach to scale gene expression data derived from different platforms for integrated analyses.', *Genomics*. Elsevier B.V., 103(4), pp. 239–51. doi: 10.1016/j.ygeno.2014.03.001.

Leek, J. T. *et al.* (2010) 'Tackling the widespread and critical impact of batch effects in high-throughput data.', *Nature reviews. Genetics*. Nature Publishing Group, 11(10), pp. 733–9. doi: 10.1038/nrg2825.

Leek, J. T. *et al.* (2012) 'The SVA package for removing batch effects and other unwanted variation in high-throughput experiments', *Bioinformatics*, 28(6), pp. 882–883. doi: 10.1093/bioinformatics/bts034.

Leek, J. T. and Storey, J. D. (2007) 'Capturing heterogeneity in gene expression studies by surrogate variable analysis', *PLoS Genetics*, 3(9), pp. 1724–1735. doi: 10.1371/journal.pgen.0030161.

Lewandowski, K. E. *et al.* (2014) 'Age as a predictor of cognitive decline in bipolar disorder.', *The American journal of geriatric psychiatry : official journal of the American Association for Geriatric Psychiatry*. NIH Public Access, 22(12), pp. 1462–8. doi: 10.1016/j.jagp.2013.10.002.

Lewis, F. *et al.* (2014) 'The Trajectory of Dementia in the UK – Making a Difference Difference', (June).

Li, J. and Tseng, G. C. (2011) 'An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies', *Annals of Applied Statistics*, 5(2 A), pp. 994–1019. doi: 10.1214/10-AOAS393.

Li, X. *et al.* (2015) 'Integrated genomic approaches identify major pathways and upstream regulators in late onset Alzheimer's disease', *Scientific Reports*. Nature Publishing Group, 5(1), p. 12393. doi: 10.1038/srep12393.

Li, Y. *et al.* (2012) 'Analysis of hippocampal gene expression profile of Alzheimer 's disease model rats using genome chip bioinformatics', 7(5), pp. 332–340. doi: 10.3969/j.issn.1673-5374. 2012.05.002.

Li, Y. *et al.* (2016) 'Implications of GABAergic Neurotransmission in Alzheimer's Disease.', *Frontiers in aging neuroscience*. Frontiers Media SA, 8, p. 31. doi: 10.3389/fnagi.2016.00031.

Liang, W. S. *et al.* (2008) 'Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons.', *Proceedings of the National Academy of Sciences of the United States of America*, 105(11), pp. 4441–6. doi: 10.1073/pnas.0709259105.

Lim, S. M., Lim, J. Y. and Cho, J. Y. (2014) 'Targeted therapy in gastric cancer: personalizing cancer

treatment based on patient genome.', *World journal of gastroenterology*. Baishideng Publishing Group Inc, 20(8), pp. 2042–50. doi: 10.3748/wjg.v20.i8.2042.

Liu, C. *et al.* (2017) 'Pathway-wide association study identifies five shared pathways associated with schizophrenia in three ancestral distinct populations', *Translational Psychiatry*, 7(2), p. e1037. doi: 10.1038/tp.2017.8.

Liu, T. *et al.* (2013) 'Transcriptional signaling pathways inversely regulated in Alzheimer's disease and glioblastoma multiform.', *Scientific Reports*, 3(1), p. 3467. doi: 10.1038/srep03467.

Loi, S. M. *et al.* (2018) 'Alzheimer disease: Non-pharmacological and pharmacological management of cognition and neuropsychiatric symptoms', *Australasian Psychiatry*. SAGE PublicationsSage UK: London, England, 26(4), pp. 358–365. doi: 10.1177/1039856218766123.

Lunnon, K. *et al.* (2012) 'Mitochondrial dysfunction and immune activation are detectable in early alzheimer's disease blood', *Journal of Alzheimer's Disease*, 30(3), pp. 685–710. doi: 10.3233/JAD-2012-111592.

Lunnon, K. *et al.* (2013) 'A blood gene expression marker of early Alzheimer's disease', *Journal of Alzheimer's Disease*, 33(3), pp. 737–753. doi: 10.3233/JAD-2012-121363.

Ma, J. *et al.* (2015) 'TYROBP in Alzheimer's Disease', *Molecular Neurobiology*, 51(2), pp. 820–826. doi: 10.1007/s12035-014-8811-9.

Maciejak, A. *et al.* (2015) 'Gene expression profiling reveals potential prognostic biomarkers associated with the progression of heart failure', *Genome Medicine*, 7(1). doi: 10.1186/s13073-015-0149-z.

Mackenzie, G. and Will, R. (2017) 'Creutzfeldt-Jakob disease: recent developments.', *F1000Research*. Faculty of 1000 Ltd, 6, p. 2053. doi: 10.12688/f1000research.12681.1.

Mackenzie, I. R. A. and Neumann, M. (2016) 'Molecular neuropathology of frontotemporal dementia: insights into disease mechanisms from postmortem studies', *Journal of Neurochemistry*. John Wiley & Sons, Ltd (10.1111), 138, pp. 54–70. doi: 10.1111/jnc.13588.

Mackenzie, I. S. *et al.* (2014) 'Incidence and prevalence of multiple sclerosis in the UK 1990-2010: a descriptive study in the General Practice Research Database.', *Journal of neurology, neurosurgery, and psychiatry*. BMJ Publishing Group Ltd, 85(1), pp. 76–84. doi: 10.1136/jnnp-2013-305450.

Maglietta, R. *et al.* (2010) 'On the reproducibility of results of pathway analysis in genome-wide expression studies of colorectal cancers', *Journal of Biomedical Informatics*. Academic Press, 43(3), pp. 397–406. doi: 10.1016/J.JBI.2009.09.005.

Maglott, D. *et al.* (2011) 'Entrez gene: Gene-centered information at NCBI', *Nucleic Acids Research*, 39(SUPPL. 1), pp. 26–31. doi: 10.1093/nar/gkq1237.

Mahley, R. W. and Rall, S. C. (2000) 'A POLIPOPROTEIN E : Far More Than a Lipid Transport Protein', (59).

Makin, S. (2018) 'The amyloid hypothesis on trial', *Nature*, 559(7715), pp. S4–S7. doi: 10.1038/d41586-018-05719-4.

Manish Pathak (2018) *Using XGBoost in Python (article) - DataCamp*. Available at: https://www.datacamp.com/community/tutorials/xgboost-in-python (Accessed: 4 March 2019).

Maouche, S. *et al.* (2008) 'Performance comparison of two microarray platforms to assess differential gene expression in human monocyte and macrophage cells', *BMC Genomics*. BioMed Central, 9(1), p. 302. doi: 10.1186/1471-2164-9-302.

Martyn, C. (2003) 'Anti-inflammatory drugs and Alzheimer's disease.', *BMJ (Clinical research ed.)*. BMJ Publishing Group, 327(7411), pp. 353–4. doi: 10.1136/bmj.327.7411.353.

Masters, C. L. *et al.* (2015) 'Alzheimer's disease', *Nature Reviews Disease Primers*. Nature Publishing Group, 1, p. 15056. doi: 10.1038/nrdp.2015.56.

Matthews, D. C., Henter, I. D. and Zarate, C. A. (2012) 'Targeting the glutamatergic system to treat Major Depressive Disorder', *Drugs*, 72(10), pp. 1313–1333. doi: 10.2165/11633130-000000000-00000.

Matthews, F. E. *et al.* (2016) 'A two decade dementia incidence comparison from the Cognitive Function and Ageing Studies I and II'. doi: 10.1038/ncomms11398.

Mattsson, N. *et al.* (2016) 'Plasma tau in Alzheimer disease.', *Neurology*. American Academy of Neurology, 87(17), pp. 1827–1835. doi: 10.1212/WNL.0000000000003246.

Mayeux, R. *et al.* (1998) 'Utility of the Apolipoprotein E Genotype in the Diagnosis of Alzheimer's Disease', *New England Journal of Medicine*, 338(8), pp. 506–511. doi: 10.1056/NEJM199802193380804.

McAleese, K. E. *et al.* (2016) 'Post-mortem assessment in vascular dementia: advances and aspirations', *BMC Medicine*. BioMed Central, 14(1), p. 129. doi: 10.1186/s12916-016-0676-5.

McArt, D. G. *et al.* (2013) 'cudaMap: a GPU accelerated program for gene expression connectivity mapping.', *BMC bioinformatics*. BioMed Central, 14, p. 305. doi: 10.1186/1471-2105-14-305.

McEvoy, L. K. *et al.* (2009) 'Alzheimer Disease: Quantitative Structural Neuroimaging for Detection and Prediction of Clinical and Structural Changes in Mild Cognitive Impairment', *Radiology*, 251(1), pp. 195–205. doi: 10.1148/radiol.2511080924.

McKhann, G. M. *et al.* (2011) 'The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease', *Alzheimer's & Dementia*, 7(3), pp. 263–269. doi: 10.1016/j.jalz.2011.03.005.

Mehta, M., Adem, A. and Sabbagh, M. (2012) 'New acetylcholinesterase inhibitors for Alzheimer's disease.', *International journal of Alzheimer's disease*. Hindawi, 2012, p. 728983. doi: 10.1155/2012/728983.

Mesko, B. *et al.* (2013) 'Peripheral blood derived gene panels predict response to infliximab in rheumatoid arthritis and Crohn's disease', *Genome Medicine*. BioMed Central Ltd, 5(6), p. 59. doi: 10.1186/gm463.

Miller, J. A. *et al.* (2013) 'Genes and pathways underlying regional and cell type changes in Alzheimer's disease', *Genome Medicine*, 5(5), p. 48. doi: 10.1186/gm452.

Miller, J. A., Oldham, M. C. and Geschwind, D. H. (2008) 'A Systems Level Analysis of Transcriptional Changes in Alzheimer's Disease and Normal Aging', *Journal of Neuroscience*, 28(6), pp. 1410–1420. doi: 10.1523/JNEUROSCI.4098-07.2008.

Mitchell, A. J. (2012) 'SensitivityPPV is a recognized test called the clinical utility index (CUI +) To cite this version ':

Mitchell, A. J. and Shiri-Feshki, M. (2009) 'Rate of progression of mild cognitive impairment to dementia - meta-analysis of 41 robust inception cohort studies', *Acta Psychiatrica Scandinavica*. John Wiley & Sons, Ltd (10.1111), 119(4), pp. 252–265. doi: 10.1111/j.1600-0447.2008.01326.x.

Miyashita, A. *et al.* (2014) 'Genes associated with the progression of neurofibrillary tangles in

alzheimer's disease', *Translational Psychiatry*, 4(March), pp. 1–8. doi: 10.1038/tp.2014.35.

Mohandas, E. and Rajmohan, V. (2009) 'Frontotemporal dementia: An updated overview.', *Indian journal of psychiatry*. Wolters Kluwer -- Medknow Publications, 51 Suppl 1(Suppl1), pp. S65-9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21416021 (Accessed: 1 February 2019).

Montagne, A., Zhao, Z. and Zlokovic, B. V (2017) 'Alzheimer's disease: A matter of blood-brain barrier dysfunction?', *The Journal of experimental medicine*. Rockefeller University Press, 214(11), pp. 3151–3169. doi: 10.1084/jem.20171406.

Montibeller, L. and de Belleroche, J. (2018) 'Amyotrophic lateral sclerosis (ALS) and Alzheimer's disease (AD) are characterised by differential activation of ER stress pathways: focus on UPR target genes', *Cell Stress and Chaperones*. Springer Netherlands, 23(5), pp. 897–912. doi: 10.1007/s12192-018-0897-y.

Montoya, A. *et al.* (2006) 'Brain imaging and cognitive dysfunctions in Huntington's disease.', *Journal of psychiatry & neuroscience : JPN*. Canadian Medical Association, 31(1), pp. 21–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16496032 (Accessed: 1 February 2019).

Mosconi, L. *et al.* (2010) 'Pre-clinical detection of Alzheimer's disease using FDG-PET, with or without amyloid imaging.', *Journal of Alzheimer's disease : JAD*. NIH Public Access, 20(3), pp. 843–54. doi: 10.3233/JAD-2010-091504.

Mukherjee, S. *et al.* (2017) 'Systems biology approach to late-onset Alzheimer's disease genome-wide association study identifies novel candidate genes validated using brain expression data and Caenorhabditis elegans experiments', *Alzheimer's & Dementia*. Elsevier Inc., (February), pp. 1–10. doi: 10.1016/j.jalz.2017.01.016.

Musa, A. *et al.* (2018) 'A review of connectivity map and computational approaches in pharmacogenomics.', *Briefings in bioinformatics*. Oxford University Press, 19(3), pp. 506–523. doi: 10.1093/bib/bbw112.

Nakamura, A. *et al.* (2018) 'High performance plasma amyloid-β biomarkers for Alzheimer's disease', *Nature*. Nature Publishing Group, 554(7691), pp. 249–254. doi: 10.1038/nature25456.

Nicole C. Berchtold, Ph.D.1, Marwan N. Sabbagh, M.D.2, Thomas G. Beach, M.D., P. D. . *et al.* (2015) 'Brain gene expression patterns differentiate Mild Cognitive Impairment from normal Aged and Alzheimer Disease', *Neurobiologyof ageing*, 27(3), pp. 320–331. doi: 10.1002/nbm.3066.Non-invasive.

Noe, E. *et al.* (2003) 'Comparison of Dementia With Lewy Bodies to Alzheimer's Disease and Parkinson's Disease With Dementia'. doi: 10.1002/mds.10633.

O'Brien, R. J. and Wong, P. C. (2011) 'Amyloid Precursor Protein Processing and Alzheimer's Disease', *Annual Review of Neuroscience*. NIH Public Access, 34(1), pp. 185–204. doi: 10.1146/annurev-neuro-061010-113613.

Ogata, H. *et al.* (1999) 'KEGG: Kyoto Encyclopedia of Genes and Genomes.', *Nucleic acids research*. Oxford University Press, 27(1), pp. 29–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9847135 (Accessed: 8 March 2019).

Ohsawa, I. *et al.* (2008) 'Age-Dependent Neurodegeneration Accompanying Memory Loss in Transgenic Mice Defective in Mitochondrial Aldehyde Dehydrogenase 2 Activity', *Journal of Neuroscience*, 28(24), pp. 6239–6249. doi: 10.1523/JNEUROSCI.4956-07.2008.

Olazarán, J. *et al.* (2015) 'A Blood-Based, 7-Metabolite Signature for the Early Diagnosis of Alzheimer's Disease', *Journal of Alzheimer's Disease*, 45, pp. 1157–1173. doi: 10.3233/JAD-142925.

Oldham, M. C., Langfelder, P. and Horvath, S. (2012) 'Network methods for describing sample relationships in genomic datasets: application to Huntington's disease.', *BMC systems biology*. BMC Systems Biology, 6(1), p. 63. doi: 10.1186/1752-0509-6-63.

Orme, T., Guerreiro, R. and Bras, J. (2018) 'The Genetics of Dementia with Lewy Bodies: Current Understanding and Future Directions.', *Current neurology and neuroscience reports*. Springer, 18(10), p. 67. doi: 10.1007/s11910-018-0874-y.

Oshiro, S., Morioka, M. S. and Kikuchi, M. (2011) 'Dysregulation of iron metabolism in Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis', *Advances in Pharmacological Sciences*, 2011. doi: 10.1155/2011/378278.

Otte, C. *et al.* (2016) 'Major depressive disorder', *Nature Reviews Disease Primers*. Nature Publishing Group, 2, p. 16065. doi: 10.1038/nrdp.2016.65.

Pannee, J. *et al.* (2014) 'The amyloid-β degradation pattern in plasma—A possible tool for clinical trials in Alzheimer's disease', *Neuroscience Letters*, 573, pp. 7–12. doi: 10.1016/j.neulet.2014.04.041.

Pannucci, C. J. and Wilkins, E. G. (2010) 'Identifying and avoiding bias in research', *Plastic and Reconstructive Surgery*, 126(2), pp. 619–625. doi: 10.1097/PRS.0b013e3181de24bc.

Paolo, G. Di and Kim, T. (2012) 'Linking Lipids to Alzheimer ' s Disease : Cholesterol and Beyond', *Aging*, 12(5), pp. 284–296. doi: 10.1038/nrn3012.Linking.

Parkkinen, J. A. and Kaski, S. (2014) 'Probabilistic drug connectivity mapping.', *BMC bioinformatics*. BioMed Central, 15, p. 113. doi: 10.1186/1471-2105-15-113.

Parsons, C. G. *et al.* (2013) 'Memantine and cholinesterase inhibitors: complementary mechanisms in the treatment of Alzheimer's disease.', *Neurotoxicity research*. Springer, 24(3), pp. 358–69. doi: 10.1007/s12640-013-9398-z.

Patel, H., Dobson, R. J. B. and Newhouse, S. J. (2019) 'A Meta-Analysis of Alzheimer's Disease Brain Transcriptomic Data', *Journal of Alzheimer's Disease*. IOS Press, Preprint(Preprint), pp. 1–22. doi: 10.3233/JAD-181085.

Patel, K. R. *et al.* (2014) 'Schizophrenia: overview and treatment options.', *P & T : a peer-reviewed journal for formulary management*. MediMedia, USA, 39(9), pp. 638–45. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25210417 (Accessed: 4 February 2019).

Patil, P. S., Aghav, J. V and Sareen, V. (no date) 'An Overview of Classification Algorithms and Ensemble Methods in Personal Credit Scoring', 7. Available at: https://archive.ics. (Accessed: 4 April 2019).

Patterson, C. (2018) 'World Alzheimer Report 2018', *THE STATE OF THE ART OF DEMENTIA RESEARCH: NEW FRONTIERS*. doi: 10.1111/j.0033-0124.1950.24_14.x.

Peters, M. J. *et al.* (2015) 'The transcriptional landscape of age in human peripheral blood.', *Nature communications*. Nature Publishing Group, 6, p. 8570. doi: 10.1038/ncomms9570.

Petersen, R. C. (2004) 'Mild cognitive impairment as a diagnostic entity', *Journal of Internal Medicine*, 256(3), pp. 183–194. doi: 10.1111/j.1365-2796.2004.01388.x.

Pietronigro, E. C. *et al.* (2017) 'NETosis in Alzheimer's disease', *Frontiers in Immunology*, 8(MAR), pp. 1–12. doi: 10.3389/fimmu.2017.00211.

Plaisier, S. B. *et al.* (2010) 'Rank-rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures.', *Nucleic acids research*. Oxford University Press, 38(17), p. e169. doi: 10.1093/nar/gkq636.

Podcasy, J. L. and Epperson, C. N. (2016) 'Considering sex and gender in Alzheimer disease and other dementias.', *Dialogues in clinical neuroscience*. Les Laboratoires Servier, 18(4), pp. 437–446. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28179815 (Accessed: 29 January 2019).

Poewe, W. *et al.* (2017) 'Parkinson disease', *Nature Reviews Disease Primers*. Nature Publishing Group, 3, p. 17013. doi: 10.1038/nrdp.2017.13.

Polliack, M. L., Barak, Y. and Achiron, A. (2001) 'Late-Onset Multiple Sclerosis', *Journal of the American Geriatrics Society*. John Wiley & Sons, Ltd (10.1111), 49(2), pp. 168–171. doi: 10.1046/j.1532-5415.2001.49038.x.

Popescu, B. F. G., Pirko, I. and Lucchinetti, C. F. (2013) 'Pathology of multiple sclerosis: where do we stand?', *Continuum (Minneapolis, Minn.)*. American Academy of Neurology, 19(4 Multiple Sclerosis), pp. 901–21. doi: 10.1212/01.CON.0000433291.23091.65.

Porcellini, E. *et al.* (2010) 'Alzheimer's disease gene signature says: beware of brain viral infections', *Immunity & Ageing*. BioMed Central Ltd, 7(1), p. 16. doi: 10.1186/1742-4933-7-16.

Preece, P. and Cairns, N. J. (2003) 'Quantifying mRNA in postmortem human brain: Influence of gender, age at death, postmortem interval, brain pH, agonal state and inter-lobe mRNA variance', *Molecular Brain Research*. Elsevier, 118(1–2), pp. 60–71. doi: 10.1016/S0169-328X(03)00337-1.

Prince, M. *et al.* (2016) 'World Alzheimer Report 2016 Improving healthcare for people living with dementia. Coverage, Quality and costs now and in the future', *Alzheimer's Disease International (ADI)*, pp. 1–140. doi: 10.13140/RG.2.2.22580.04483.

Puthiyedth, N. *et al.* (2016) 'Identification of differentially expressed genes through integrated study of Alzheimer's disease affected brain regions', *PLoS ONE*. Public Library of Science, 11(4), pp. 1–29. doi: 10.1371/journal.pone.0152342.

Ramanan, V. K. *et al.* (2013) 'Genome-wide pathway analysis of memory impairment in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort implicates gene candidates, canonical pathways, and networks', 6(4), pp. 634–648. doi: 10.1007/s11682-012-9196-x.Genome-wide.

Rao, J. S. *et al.* (2012) 'Epigenetic modifications in frontal cortex from Alzheimer's disease and bipolar disorder patients', *Translational Psychiatry*, 2(7), pp. e132–e132. doi: 10.1038/tp.2012.55.

Readhead, B. *et al.* (2018) 'Multiscale Analysis of Independent Alzheimer's Cohorts Finds Disruption of Molecular, Genetic, and Clinical Networks by Human Herpesvirus', *Neuron*. Elsevier Inc., 99(1), pp. 64-82.e7. doi: 10.1016/j.neuron.2018.05.023.

Reilly, S. *et al.* (2012) 'The Role of Primary Care in Service Provision for People with Severe Mental Illness in the United Kingdom', *PLoS ONE*. Edited by H. R. Baradaran, 7(5), p. e36468. doi: 10.1371/journal.pone.0036468.

Reiman, E. M. *et al.* (2012) 'Brain imaging and fluid biomarker analysis in young adults at genetic risk for autosomal dominant Alzheimer's disease in the presenilin 1 E280A kindred: a case-control study.', *The Lancet. Neurology*. NIH Public Access, 11(12), pp. 1048–56. doi: 10.1016/S1474-4422(12)70228-4.

Reiner, A., Dragatsis, I. and Dietrich, P. (2011) 'Genetics and neuropathology of Huntington's disease.', *International review of neurobiology*. NIH Public Access, 98, pp. 325–72. doi: 10.1016/B978-0-12-381328-2.00014-6.

Richard, A. C. *et al.* (2014) 'Comparison of gene expression microarray data with count-based RNA measurements informs microarray interpretation', *BMC Genomics*. BioMed Central Ltd., 15(1). doi: 10.1186/1471-2164-15-649.

Ringnér, M. (2008) *What is principal component analysis?*, *NATURE BIOTECHNOLOGY*. Available at: http://www.nature.com/naturebiotechnology (Accessed: 26 February 2019).

Roberts, R. and Knopman, D. S. (2013) 'Classification and epidemiology of MCI.', *Clinics in geriatric medicine*. NIH Public Access, 29(4), pp. 753–72. doi: 10.1016/j.cger.2013.07.003.

Rockwood, K. *et al.* (2007) 'The clinical meaningfulness of ADAS-Cog changes in Alzheimer's disease patients treated with donepezil in an open-label trial.', *BMC neurology*. BioMed Central, 7, p. 26. doi: 10.1186/1471-2377-7-26.

Roeben, B. *et al.* (2016) 'Association of Plasma Aβ40 Peptides, But Not Aβ42, with Coronary Artery Disease and Diabetes Mellitus', *Journal of Alzheimer's Disease*, 52(1), pp. 161–169. doi: 10.3233/JAD-150575.

Roed, L. *et al.* (2013) 'Prediction of mild cognitive impairment that evolves into alzheimer's disease dementia within two years using a gene expression signature in blood: A pilot study', *Journal of Alzheimer's Disease*, 35(3), pp. 611–621. doi: 10.3233/JAD-122404.

Rong, X.-F. *et al.* (2017) 'The pathological roles of NDRG2 in Alzheimer's disease, a study using animal models and APPwt-overexpressed cells', *CNS Neuroscience & Therapeutics*, 23(8), pp. 667–679. doi: 10.1111/cns.12716.

Rosen, W. G., Mohs, R. C. and Davis, K. L. (1984) 'A new rating scale for Alzheimer's disease', *American Journal of Psychiatry*, 141(11), pp. 1356–1364. doi: 10.1176/ajp.141.11.1356.

Rouchka, E. C., Phatak, A. W. and Singh, A. V. (2008) 'Effect of single nucleotide polymorphisms on Affymetrix match-mismatch probe pairs.', *Bioinformation*. Biomedical Informatics Publishing Group, 2(9), pp. 405–11. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18795114 (Accessed: 25 February 2019).

Rumsey, D. J. (Deborah J. (2011) *Statistics for dummies*. Wiley. Available at: https://www.dummies.com/store/product/Statistics-For-Dummies-2nd-Edition.productCd-1119293529.html (Accessed: 7 March 2019).

Rye, P. D. *et al.* (2011) 'A novel blood test for the early detection of Alzheimer's disease', *Journal of Alzheimer's Disease*, 23(1), pp. 121–129. doi: 10.3233/JAD-2010-101521.

Sabbagh, M. N. *et al.* (2017) 'Increasing Precision of Clinical Diagnosis of Alzheimer's Disease Using a Combined Algorithm Incorporating Clinical and Novel Biomarker Data.', *Neurology and therapy*. Springer, 6(Suppl 1), pp. 83–95. doi: 10.1007/s40120-017-0069-5.

Saberi, S. *et al.* (2015) 'Neuropathology of Amyotrophic Lateral Sclerosis and Its Variants.', *Neurologic clinics*. NIH Public Access, 33(4), pp. 855–76. doi: 10.1016/j.ncl.2015.07.012.

Safari-Alighiarloo, N. *et al.* (2014) 'Protein-protein interaction networks (PPI) and complex diseases.', *Gastroenterology and hepatology from bed to bench*. Shahid Beheshti University of Medical Sciences, 7(1), pp. 17–31. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25436094 (Accessed: 4 March 2019).

Salzberg, S. L. (2018) 'Open questions: How many genes do we have?', *BMC Biology*. BioMed Central, 16(1), p. 94. doi: 10.1186/s12915-018-0564-x.

Saura, C. A., Parra-Damas, A. and Enriquez-Barreto, L. (2015) 'Gene expression parallels synaptic excitability and plasticity changes in Alzheimer's disease', *Frontiers in Cellular Neuroscience*, 9(August), pp. 1–14. doi: 10.3389/fncel.2015.00318.

Savica, R. *et al.* (2013) 'Incidence of Dementia With Lewy Bodies and Parkinson Disease Dementia', *JAMA Neurology*, 70(11), p. 1396. doi: 10.1001/jamaneurol.2013.3579.

Scheltens, P. *et al.* (2002) 'Structural magnetic resonance imaging in the practical assessment of dementia: beyond exclusion', *The Lancet Neurology*. Elsevier, 1(1), pp. 13–21. doi: 10.1016/S1474-4422(02)00002-9.

Scherzer, C. R. *et al.* (2006) 'Molecular markers of early Parkinson ' s disease based on gene expression in blood'.

Schizophrenia - Symptoms - NHS (2016). Available at: https://www.nhs.uk/conditions/schizophrenia/symptoms/ (Accessed: 4 February 2019).

Schmid, R. *et al.* (2010) 'Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3', *BMC Genomics*. BioMed Central, 11(1), p. 349. doi: 10.1186/1471-2164-11-349.

Schousboe, A. *et al.* (2014) 'Glutamate and ATP at the Interface of Metabolism and Signaling in the Brain', 11, pp. 13–30. doi: 10.1007/978-3-319-08894-5.

Sekar, S. *et al.* (2015) 'Alzheimer's disease is associated with altered expression of genes involved in immune response and mitochondrial processes in astrocytes', *Neurobiology of Aging*. Elsevier Inc, 36(2), pp. 583–591. doi: 10.1016/j.neurobiolaging.2014.09.027.

Sellam, J. *et al.* (2014) 'Use of whole-blood transcriptomic profiling to highlight several pathophysiologic pathways associated with response to rituximab in patients with rheumatoid arthritis: Data from a randomized, controlled, open-label trial', *Arthritis and Rheumatology*, 66(8), pp. 2015–2025. doi: 10.1002/art.38671.

Serrano-Pozo, A. *et al.* (2011) 'Neuropathological alterations in Alzheimer disease.', *Cold Spring Harbor perspectives in medicine*, 1(1), pp. 1–23. doi: 10.1101/cshperspect.a006189.

Sethi, A. J. *et al.* (2016) 'Microbes and Alzheimer's Disease', 335(6068), pp. 590–593. doi: 10.1126/science.1212867.Sequential.

Sevigny, J. *et al.* (2016) 'Amyloid PET screening for enrichment of early-stage Alzheimer disease clinical trials: Experience in a phase 1b clinical trial', *Alzheimer Disease and Associated Disorders*, 30(1), pp. 1–7. doi: 10.1097/WAD.0000000000000144.

Shoffner, J. M. (1997) 'Oxidative phosphorylation defects and Alzheimer's disease', *Neurogenetics*, 1(1), pp. 13–19. doi: 10.1007/s100480050002.

Siavelis, J. C. *et al.* (2016a) 'Bioinformatics methods in drug repurposing for Alzheimer's disease', *Briefings in Bioinformatics*. BioMed Central Ltd, 17(2), pp. 322–335. doi: 10.1093/bib/bbv048.

Siavelis, J. C. *et al.* (2016b) 'Bioinformatics methods in drug repurposing for Alzheimer's disease', *Briefings in Bioinformatics*, 17(2), pp. 322–335. doi: 10.1093/bib/bbv048.

Silver, J. D., Ritchie, M. E. and Smyth, G. K. (2009) 'Microarray background correction: maximum likelihood estimation for the normal-exponential convolution.', *Biostatistics (Oxford, England)*. Oxford University Press, 10(2), pp. 352–63. doi: 10.1093/biostatistics/kxn042.

Šimundić, A.-M. (2009) 'Measures of Diagnostic Accuracy: Basic DefinitiŠimundić, A.-M. (2009). Measures of Diagnostic Accuracy: Basic Definitions. EJIFCC, 19(4), 203–11. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27683318ons.', *Ejifcc*, 19(4), pp. 203–11. doi: eJIFCC2008Vol19No4pp203-211.

Šimundić, A. M. (2013) 'Bias in research', *Biochemia Medica*, 23(1), pp. 12–15. doi: 10.11613/BM.2013.003.

Singh, D. *et al.* (2014) 'Altered gene expression in blood and sputum in copd frequent exacerbators in the eclipse cohort', *PLoS ONE*, 9(9), pp. 1–9. doi: 10.1371/journal.pone.0107381.

Sinnaeve, P. R. *et al.* (2009) 'Gene expression patterns in peripheral blood correlate with the extent of coronary artery disease.', *PloS one*, 4(9), p. e7037. doi: 10.1371/journal.pone.0007037.

Sîrbu, A. *et al.* (2012) 'RNA-Seq vs dual- and single-channel microarray data: sensitivity analysis for differential expression and clustering.', *PloS one*, 7(12), p. e50986. doi: 10.1371/journal.pone.0050986.

Smyth, G. (2005) 'limma: Linear Models for Microarray Data', *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, (2005), pp. 397–420. doi: citeulike-article-id:5722720.

Del Sole, A., Malaspina, S. and Magenta Biasina, A. (2016) 'Magnetic resonance imaging and positron emission tomography in the diagnosis of neurodegenerative dementias.', *Functional neurology*. CIC Edizioni Internazionali, 31(4), pp. 205–215. doi: 10.11138/FNEUR/2016.31.4.205.

Sood, S. *et al.* (2015) 'A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status.', *Genome biology*. Genome Biology, 16, p. 185. doi: 10.1186/s13059-015-0750-x.

Sperling, R. A. *et al.* (2011) 'Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease', *Alzheimer's & Dementia*, 7(3), pp. 280–292. doi: 10.1016/j.jalz.2011.03.003.

Stern, Y. (2002) 'What is cognitive reserve? Theory and research application of the reserve concept.', *Journal of the International Neuropsychological Society : JINS*, 8(3), pp. 448–60. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11939702 (Accessed: 2 January 2019).

Storandt, M. and Morris, J. C. (2010) 'Ascertainment bias in the clinical diagnosis of Alzheimer disease', *Archives of Neurology*, 67(11), pp. 1364–1369. doi: 10.1001/archneurol.2010.272.

Strong, M. J., Kesavapany, S. and Pant, H. C. (2005) 'The Pathobiology of Amyotrophic Lateral Sclerosis: A Proteinopathy?', *Journal of Neuropathology and Experimental Neurology*. Oxford University Press, 64(8), pp. 649–664. doi: 10.1097/01.jnen.0000173889.71434.ea.

Subramanian, A. *et al.* (2005) 'Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 102(43), pp. 15545–50. doi: 10.1073/pnas.0506580102.

Subramanian, A. *et al.* (2017) 'A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles.', *Cell*. NIH Public Access, 171(6), pp. 1437-1452.e17. doi: 10.1016/j.cell.2017.10.049.

Swerdlow, R. H. and Khan, S. M. (2004) 'A "mitochondrial cascade hypothesis" for sporadic Alzheimer's disease', *Medical Hypotheses*. Churchill Livingstone, 63(1), pp. 8–20. doi: 10.1016/J.MEHY.2003.12.045.

Tabatabaei-Jafari, H., Shaw, M. E. and Cherbuin, N. (2015) 'Cerebral atrophy in mild cognitive impairment: A systematic review with meta-analysis', *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*. Elsevier Inc., 1(4), pp. 487–504. doi: 10.1016/j.dadm.2015.11.002.

Takeda, T. (2018) 'Possible concurrence of TDP-43, tau and other proteins in amyotrophic lateral sclerosis/frontotemporal lobar degeneration', *Neuropathology*. John Wiley & Sons, Ltd (10.1111), 38(1), pp. 72–81. doi: 10.1111/neup.12428.

Taminau, J. *et al.* (2012) 'Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages', *BMC Bioinformatics*, 13, p. 335. doi:

10.1186/1471-2105-13-335.

Tarca, A. L. *et al.* (2009) 'A novel signaling pathway impact analysis', *Bioinformatics*, 25(1), pp. 75–82. doi: 10.1093/bioinformatics/btn577.

Thalhauser, C. J. and Komarova, N. L. (2012) 'Alzheimer's disease: rapid and slow progression.', *Journal of the Royal Society, Interface*. The Royal Society, 9(66), pp. 119–26. doi: 10.1098/rsif.2011.0134.

Thambisetty, M. and Lovestone, S. (2010) 'Blood-based biomarkers of Alzheimer's disease: challenging but feasible.', *Biomarkers in medicine*, 4(1), pp. 65–79. doi: 10.2217/bmm.09.84.

The ALS Association (no date). Available at: http://www.alsa.org/als-care/resources/publications-videos/factsheets/epidemiology.html (Accessed: 31 January 2019).

Tušar, T. *et al.* (2017) 'A study of overfitting in optimization of a manufacturing quality control procedure', *Applied Soft Computing*. Elsevier, 59, pp. 77–87. doi: 10.1016/J.ASOC.2017.05.027.

Vawter, M. P. *et al.* (2004) 'Gender-specific Gene Expression in Post-Mortem Human Barin: Localization to Sex Chromosomes', *Neuropsychopharmacology*, 29(2), pp. 373–384. doi: 10.1038/sj.npp.1300337.Gender-Specific.

Venet, D., Dumont, J. E. and Detours, V. (2011) 'Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome', *PLoS Computational Biology*. Edited by I. Rigoutsos. Public Library of Science, 7(10), p. e1002240. doi: 10.1371/journal.pcbi.1002240.

Vieta, E. *et al.* (2018) 'Bipolar disorders', *Nature Reviews Disease Primers*. Nature Publishing Group, 4, p. 18008. doi: 10.1038/nrdp.2018.8.

Villemagne, V. L. *et al.* (2014) 'In vivo evaluation of a novel tau imaging tracer for Alzheimer's disease', *European Journal of Nuclear Medicine and Molecular Imaging*. Springer Berlin Heidelberg, 41(5), pp. 816–826. doi: 10.1007/s00259-013-2681-7.

Viña, J., Viña, V. and Lloret, A. (2010) 'Why Women Have More Alzheimer's Disease Than Men: Gender and Mitochondrial Toxicity of Amyloid-β Peptide', *Journal of Alzheimer's Disease*, 20, pp. 527–533. doi: 10.3233/JAD-2010-100501.

Voyle, N. *et al.* (2016) 'A pathway based classification method for analyzing gene expression for Alzheimer's disease diagnosis', *Journal of Alzheimer's Disease*, 49(3), pp. 659–669. doi: 10.3233/JAD-150440.

Walker, Z. *et al.* (2012) 'Comparison of cognitive decline between dementia with Lewy bodies and Alzheimer's disease: a cohort study.', *BMJ open*. British Medical Journal Publishing Group, 2(1), p. e000380. doi: 10.1136/bmjopen-2011-000380.

Walsh, A. M. *et al.* (2016) 'Integrative genomic deconvolution of rheumatoid arthritis GWAS loci into gene and cell type associations', *Genome biology*. Genome Biology, 17, p. 79. doi: 10.1186/s13059-016-0948-6.

Walton, H. S. and Dodd, P. R. (2007) 'Glutamate-glutamine cycling in Alzheimer's disease', *Neurochemistry International*, 50(7–8), pp. 1052–1066. doi: 10.1016/j.neuint.2006.10.007.

Wang, H.-X., Xu, W. and Pei, J.-J. (2012) 'Leisure activities, cognition and dementia', *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. Elsevier, 1822(3), pp. 482–491. doi: 10.1016/J.BBADIS.2011.09.002.

Wang, Z., Gerstein, M. and Snyder, M. (2009) 'RNA-Seq: a revolutionary tool for transcriptomics.', *Nature reviews. Genetics*. Nature Publishing Group, 10(1), pp. 57–63. doi: 10.1038/nrg2484.

Weiner, H. L. and Selkoe, D. J. (2002) 'Inflammation and therapeutic vaccination in CNS diseases', *Nature*, 420(6917), pp. 879–884. doi: 10.1038/nature01325.

Williams, G. (2012) 'A searchable cross-platform gene expression database reveals connections between drug treatments and disease.', *BMC genomics*. BioMed Central Ltd, 13(1), p. 12. doi: 10.1186/1471-2164-13-12.

Wu, Z. *et al.* (2004) 'A Model-Based Background Adjustment for Oligonucleotide Expression Arrays'. doi: 10.1198/016214504000000683.

Xia, J., Benner, M. J. and Hancock, R. E. W. (2014) 'NetworkAnalyst - Integrative approaches for protein-protein interaction network analysis and visual exploration', *Nucleic Acids Research*, 42(W1), pp. 167–174. doi: 10.1093/nar/gku443.

Xue, H. *et al.* (2018) 'Review of Drug Repositioning Approaches and Resources.', *International journal of biological sciences*. Ivyspring International Publisher, 14(10), pp. 1232–1244. doi: 10.7150/ijbs.24612.

Yang, H. *et al.* (2016) 'Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information', *Nucleic Acids Research*. Oxford University Press, 44(D1), pp. D1069–D1074. doi: 10.1093/nar/gkv1230.

Yang, H. *et al.* (2019) 'AdmetSAR 2.0: Web-service for prediction and optimization of chemical ADMET properties', *Bioinformatics*. Oxford University Press, 35(6), pp. 1067–1069. doi: 10.1093/bioinformatics/bty707.

Yella, J. K. *et al.* (2018) 'Changing Trends in Computational Drug Repositioning.', *Pharmaceuticals (Basel, Switzerland)*. Multidisciplinary Digital Publishing Institute (MDPI), 11(2). doi: 10.3390/ph11020057.

Yoshiike, Y. *et al.* (2008) 'GABA(A) receptor-mediated acceleration of aging-associated memory decline in APP/PS1 mice and its pharmacological treatment by picrotoxin.', *PloS one*. Public Library of Science, 3(8), p. e3029. doi: 10.1371/journal.pone.0003029.

Yvette, G. *et al.* (2010) 'Beta-Amyloid Deposition and the Aging Brain Karen', *North*, 11(4), pp. 436–450. doi: 10.1007/s11065-009-9118-x.Beta-Amyloid.

Zaki, N. and Mora, A. (2015) 'A comparative analysis of computational approaches and algorithms for protein subcomplex identification', *Scientific Reports*, 4(1), p. 4262. doi: 10.1038/srep04262.

Zarei, S. *et al.* (2015) 'A comprehensive review of amyotrophic lateral sclerosis.', *Surgical neurology international*. Wolters Kluwer -- Medknow Publications, 6, p. 171. doi: 10.4103/2152-7806.169561.

Zhang, M. *et al.* (2008) 'Apparently low reproducibility of true differential expression discoveries in microarray studies', *Bioinformatics*, 24(18), pp. 2057–2063. doi: 10.1093/bioinformatics/btn365.

Zhang, S.-D. and Gant, T. W. (2008) 'A simple and robust method for connecting small-molecule drugs using gene-expression signatures.', *BMC bioinformatics*. BioMed Central, 9, p. 258. doi: 10.1186/1471-2105-9-258.

Zhao, S. *et al.* (2014) 'Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells', *PLoS ONE*. Edited by S.-D. Zhang. Public Library of Science, 9(1), p. e78644. doi: 10.1371/journal.pone.0078644.

Zhou, Z.-H. (Computer scientist) (2012) *Ensemble methods : foundations and algorithms*. Taylor & Francis. Available at: https://www.crcpress.com/Ensemble-Methods-Foundations-and-Algorithms/Zhou/p/book/9781439830031 (Accessed: 1 April 2019).

Zhu, X. *et al.* (2004) 'Alzheimer's disease: the two-hit hypothesis', *The Lancet Neurology*. Elsevier, 3(4), pp. 219–226. doi: 10.1016/S1474-4422(04)00707-0.

Zlomuzica, A. *et al.* (2016) 'Neuronal histamine and cognitive symptoms in Alzheimer's disease', *Neuropharmacology*. Pergamon, 106, pp. 135–145. doi: 10.1016/J.NEUROPHARM.2015.05.007.

Zubenko, G. S. *et al.* (2014) 'Differential hippocampal gene expression and pathway analysis in an etiology-based mouse model of major depressive disorder', *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, 165(6), pp. 457–466. doi: 10.1002/ajmg.b.32257.

Appendices

Appendix A

Chapter 2 Supplementary Material

Supplementary Table 2.1: Significance testing of gender, age and PM delay between diagnosis groups

Tissue	Comparison	Gender	Age	PM Delay
		P-value	P-value	P-value
Cerebellum	AD vs AsymAD	0.25	0.00	0.03
	AD vs CO	0.77	0.11	0.01
	AsymAD vs CO	0.15	0.00	0.55
Entorhinal Cortex	AD vs AsymAD	0.60	0.00	0.01
	AD vs CO	0.37	0.38	0.04
	AsymAD vs CO	0.14	0.01	0.22
Frontal Cortex	AD vs AsymAD	0.99	0.00	0.12
	AD vs CO	0.15	0.06	0.01
	AsymAD vs CO	0.11	0.00	0.86
Temporal Cortex	AD vs AsymAD	0.42	0.00	0.03
	AD vs CO	0.39	0.05	0.02
	AsymAD vs CO	0.11	0.01	0.45

The table provides the Mann Whitney U test results which suggest Age is significantly lower in the control groups. "AsymAD vs CO" represents a comparison between the control and AsymAD group. "AD vs AsymAD" represents a comparison between the AsymAD and AD group. "AD vs CO" represents a comparison between the control and AD group. Abbreviation: PM = Post Mortem

Supplementary Table 2.2: Differentially expression analysis results when accounting for age

Tissue	Comparison	N.o. probes	N.o. controls	N.o. probes	N.o. significant DEG	N.o. significant DEG (with age)
Entorhinal Cortex	AD vs CO	34	16	3518	1690	1435
	AD vs AsymAD	34	28	3518	1904	1942
	AsymAD vs CO	28	16	3518	19	1
Temporal Cortex	AD vs CO	45	24	3518	1517	1383
	AD vs AsymAD	45	28	3518	1546	1532
	AsymAD vs CO	28	24	3518	253	56
Frontal Cortex	AD vs CO	38	21	3518	299	195
	AD vs AsymAD	38	32	3518	52	40
	CO_AD vs CO	**32**	**21**	**3518**	**398**	**436**
Cerebellum	**AD vs CO**	**36**	**18**	**3518**	**176**	**76**
	AD vs CO_AD	**36**	**27**	**3518**	**13**	**13**
	CO_AD vs CO	**27**	**18**	**3518**	**1**	**1**

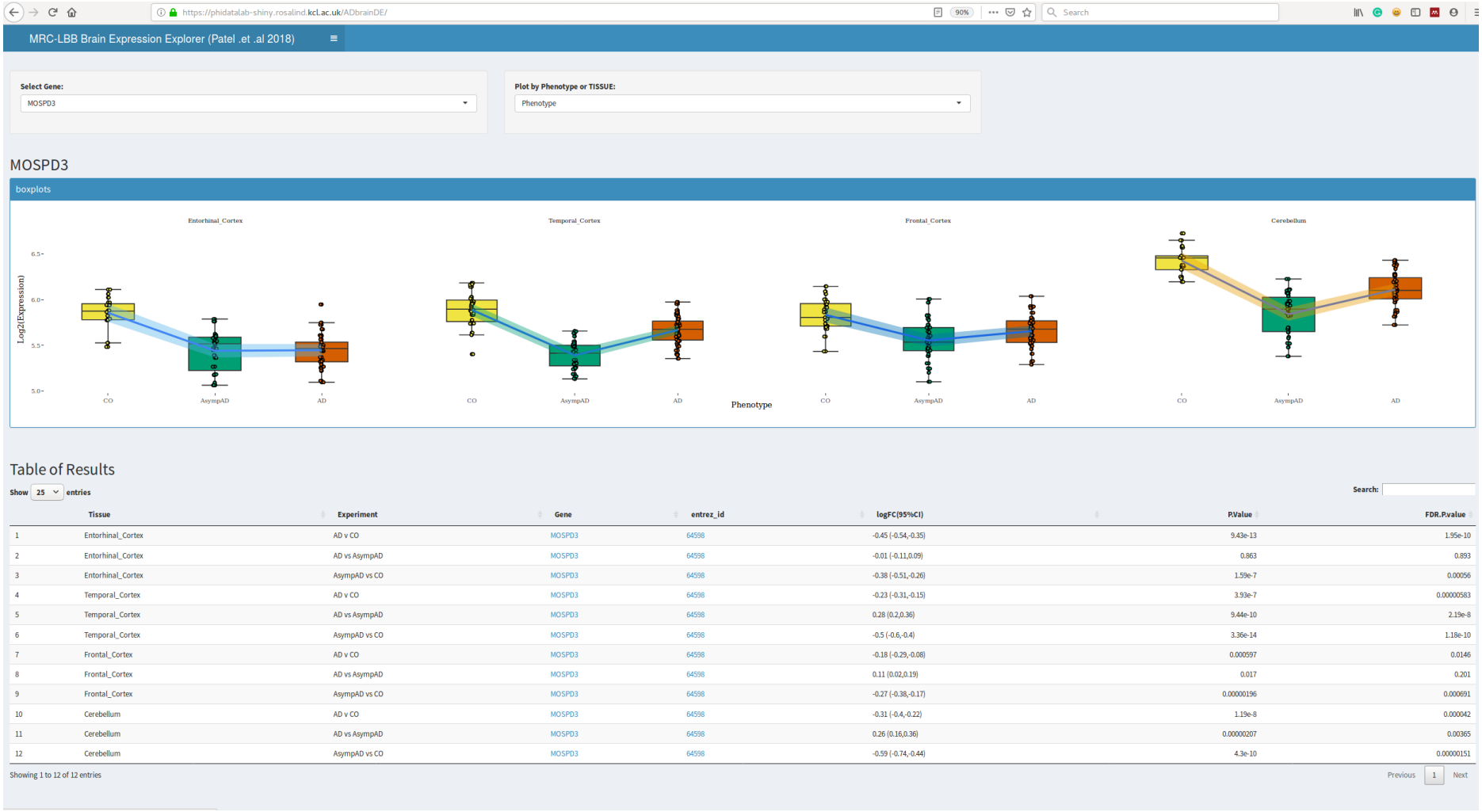
The table shows the effects of incorporating age into the model when performing differential expression analysis. In general, the number of differentially expressed genes is reduced when accounting for age. A gene is regarded to be significantly differentially expressed when the adjusted p-value is ≤ 0.05. "AsymAD vs CO" represents a comparison between the control and AsymAD group. "AD vs AsymAD" represents a comparison between the AsymAD and AD group. "AD vs CO" represents a comparison between the control and AD group. Abbreviation: PM = DEG: Differential Expressed Gene

6.5.1 MRC-LBB brain expression explorer

The "MRC-LBB brain expression explorer" R shiny web application provides gene-level results to the broader research community. The application was developed in R using the "shiny" framework (version 0.14) and is hosted on the KCL Rosalind OpenStack service (http://rosalind.kcl.ac.uk/). The application can be accessed at https://phidatalab-shiny.rosalind.kcl.ac.uk/ADbrainDE/, and the raw code is accessible at https://doi.org/10.5281/zenodo.1400644.

The application allows users to query a gene of interest and observe the change in expression of this gene between the three phenotype groups (cognitive healthy controls, AsymAD and AD subjects) and across the four different brain regions (entorhinal cortex, temporal cortex, frontal cortex and cerebellum). The differential expression results are also provided for the gene of interest, providing the user with the ability to determine if the gene is significantly differentially expressed and how the gene is regulated between the phenotype groups. Furthermore, this information is graphically displayed in the form of boxplots to help the user explore and understand the expression changes further. A screenshot of the application is provided in Supplementary Figure 2.1

To use the application, users can select any one of the 3518 genes deemed as reliably detected in the study from the "Select Gene" drop-down menu. Selecting a gene will automatically populate the boxplots and table of differential expression results. Users can use the "Plot by Phenotype or TISSUE" drop-down menu to display results grouped by brain regions of by phenotype. The table of results will display twelve rows of result, each being a different analysis of brain region by phenotype. The table provides the differential expression logFC and associated p-value and FDR adjusted p-value for the selected gene.



Supplementary Figure 2.1: Screenshot of the "MRC-LBB brain expression explorer" R shiny web application

Appendix B

Chapter 4 Supplementary Material

Supplementary Table 4.1: Count of CMap compounds approved to treat diseases

Disease	N.o. treatments in the CMap database
Bacterial infection	66
Hypertension	42
Pain	23
Schizophrenia	22
Depression	21
Fungal infection	15
Diabetes mellitus	14
Angina	11
Rheumatoid arthritis	11
Anaesthesia	10
Parkinson's disease	10
Allergic rhinitis	9
Breast cancer	9
Inflammatory diseases	9
Malaria	8
Nausea	8
Anxiety disorder	7
Cancers	7
Congestive heart failure	7
Epilepsy	7
High blood pressure	7
Osteoarthritis	7
Urinary tract infections	7
Vomiting	7
Asthma	6
Bipolar disorder	6
Acne vulgaris	5
Arthritis	5
Cancer	5
Cardiovascular disease	5
Oedema	5
Gout	5
HIV infections	5
Tuberculosis	5
Ventricular arrhythmias	5
Alcohol use disorders	4
Allergy	4
Chronic obstructive pulmonary disease	4
Dysmenorrhea	4
Oedema associated with congestive heart failure	4
Glaucoma	4
Major depressive disorder	4
Multiple sclerosis	4

Open-angle glaucoma	4
Peptic ulcer disease	4
Prostate cancer	4
Psychotic disorders	4
Worm infections	4
Alzheimer's disease	3
Benign prostatic hyperplasia	3
Cough	3
Cushing's disease	3
Heart failure	3
Hormone replacement therapy	3
Hypercholesterolemia	3
Hyperlipidemia	3
Musculoskeletal disease	3
Obesity	3
Tachyarrhythmias	3
Viral infections	3
Acid-reflux disorders	2
Acute promyelocytic leukaemia	2
Allergic conjunctivitis	2
Angina pectoris	2
Ankylosing spondylitis	2
Atrial fibrillation	2
Brain cancer	2
Cancer pain	2
Candida infection	2
Cerebrovascular disorders	2
Cognitive impairment	2
Colorectal cancer	2
Constipation	2
Cutaneous T-cell lymphoma	2
Dermatitis	2
Diarrhoea	2
Erectile dysfunction	2
Gastroesophageal reflux disease	2
Hay fever	2
Hepatitis C virus infection	2
Herpes simplex virus infection	2
Hyperthyroidism	2
Hypotension	2
Infectious disease	2
Melanoma	2
Migraine	2
Multiple myeloma	2
Organ transplant rejection	2
Osteoporosis	2

Parasitic infection	2
Premature labour	2
Pseudomonas infection	2
Psoriasis	2
Pulmonary arterial hypertension	2
Skin allergies	2
Vertigo Meniere's disease	2
Vitiligo	2
Vulnerary	2
Wound healing	2
Acquired nystagmus	1
Active rheumatoid arthritis	1
Acute and chronic intestinal amebiasis	1
Acute asthma	1
Acute diarrhoea	1
Acute gouty arthritis	1
Acute hepatic failure and alcoholic cirrhosis of the liver	1
Acute ischemic stroke	1
Acute malaria	1
Acute myeloid leukaemia	1
Acute pain	1
Addiction	1
Advanced breast and endometrium carcinoma	1
Advanced gum disease	1
Alcoholic hepatitis	1
Allergic disorders	1
Amebiasis	1
Amoebiasis	1
Amyotrophic lateral sclerosis	1
Analgesic	1
Anaemia	1
Antispasmodic and uricosuric	1
Arrhythmias	1
Arthropathy	1
Atherosclerosis	1
Athlete's foot	1
Atrophic vaginitis	1
Attention deficit hyperactivity disorder	1
Bacterial respiratory tract infection	1
Bladder disease	1
Blood clotting disorder	1
Brain ischaemia	1
Bronchial asthma	1
Bronchitis	1
Bronchodilator	1
Capillary fragility	1

Cardiac arrhythmias	1
Cardiogenic shock	1
Central nervous system disease	1
Cerebral infarction	1
Cerebral salt-wasting syndrome	1
Cerebral stimulant	1
Cerebral vasospasm	1
Cerebrovascular ischaemia	1
Cestode infection	1
Chemotherapy-induced nausea and vomiting	1
Chemotherapy or radiotherapy-induced mucositis	1
Chronic breathing disorders	1
Chronic lymphocytic leukaemia	1
Chronic pain	1
chronic periodontitis	1
Chronic stable angina	1
Cocaine addiction	1
Cognitive disorders	1
Colon cancer	1
Complex partial seizures	1
Condyloma	1
Congestive cardiac insufficiency	1
Conjunctivitis	1
Coronary artery disease	1
Coronary heart disease	1
Crohn's disease	1
Cystic fibrosis	1
Dementia	1
Dermal necrosis	1
Dermatitis herpetiformis	1
Dermatological disease	1
Dermatomycosis	1
Diagnosis of mydriasis	1
Dutch elm disease	1
Dyspepsia	1
Emesis	1
Endocarditis	1
Epileptic conditions	1
Excessive bleeding	1
Eye Inflammation	1
Female genital tract inflammation	1
Female infertility	1
Flatworms infection	1
Folate deficiency	1
Fungal urinary tract infection	1
Gastric secretory disorders	1

Gastrointestinal disorders	1
Gastrointestinal problems	1
Glioma	1
Gonorrheal vaginitis	1
Gram-negative bacterial infection	1
Hematologic malignancies	1
Hepatitis virus infection	1
High cholesterol levels	2
High triglyceride levels	1
Hodgkin's lymphoma	1
Hormonal contraceptives	1
Hormonally-responsive breast cancer	1
Hormone deficiency	1
Hyperlipidaemia	1
Hyperlipoproteinemia	1
hyperuricemia	1
Hypogonadism	1
Hypoparathyroidism	1
Hypothyroidism	1
Inflammation and itching	1
Inflammatory bowel disease	1
Influenza virus A infection	1
Insecticide	1
Insomnia	1
Intermittent claudication	1
Intestinal strongyloidiasis due to nematode parasite	1
Irregularities	1
Irritable bowel syndrome	1
Itching	1
Jock itch	1
Joint and muscular pain	1
Kidney cancer	1
Fluid retention	1
Lymphatic filariasis	1
Maintenance of normal sinus rhythm	1
Male impotence	1
Malignant hyperthermia	1
Malignant melanoma	1
Medical abortion	1
Menopausal and postmenopausal disorders	1
Menorrhagia	1
Menstrual disorders	1
Metastatic breast cancer	1
Metastatic islet cell carcinoma	1
Migraine headaches	1
Minor eye irritations	1

Miosis during ocular surgery	1
Morning sickness	1
MRSA infection	1
Muscle Relaxant	1
Muscle spasm	1
Mycobacterium avium complex disease	1
Mycobacterium tuberculosis infection	1
Myelodysplastic Syndromes	1
Narcotic depression	1
Nasal congestion	1
Nausea and vomiting	1
Neuropathic pain	1
Non-infectious rhinitis	1
Non-insulin dependent diabetes	1
NSAID-related gastric ulcer	1
Ocular disease	1
Ocular hypertension	1
Oedema and hypertension	1
Oral contraceptives	1
Organophosphate poisoning	1
Organ rejection	1
Orthostatic hypotension	1
Otitis media	1
Ovarian cancer	1
Paget's disease	1
Pancreatic cancer	1
Pancreatic disorders	1
Pancreatitis	1
Partial seizures	1
Pediatric eye examinations	1
Peripheral nerve damage	1
Peripheral vascular disease	1
Peripheral vasoconstriction	1
Photosensitizer	1
Piscicide	1
Pneumocystis carinii pneumonia	1
Poison intoxication	1
Prostate tumour	1
Pruritus	1
Psychomotor agitation	1
Psychoses	1
Pulmonary embolism	1
Pulmonary hypertension	1
Rheumatoid arthritis	1
Ringworm infections	1
Schistosomiasis	1

Seizures	1
Senile dementia	1
Skin infections	1
Skin photodamage	1
Sleep disorders	1
Solid tumours	1
Spasm	1
Spasms	1
Spinal anaesthesia	1
Spinal cord injuries	1
Stabilize muscle contractions	1
Stroke	1
Substance dependence	1
Thromboembolic disorders	1
Thrombosis	1
Tinea pedis	1
Trypanosomiasis	1
Ulcerative colitis	1
Urinary incontinence	1
Urinary retention	1
Vascular disease	1
Vasodilator, diuretic	1
Vitamin deficiency	1
Xerophthalmia	1
Xerostomia	1
X-rays imaging	1

The TTD contained information on 31,614 drug-disease relations and were used to identify diseases which can be treated by any of the 1146 CMap compounds. The CMap database contains 1121 unique compounds, with some compounds experimentally repeated at different concentrations and therefore, generated 1146 expression profiles. Only 612 CMap compounds annotated to 299 different diseases; however, due to multiple drug-disease combinations, 768 CMap drug-disease relations were identified and are provided above.

Appendix C

Chapter 5 Supplementary Material

Supplementary Table 5.1: "AD vs mixed control" Classification model's predictive genes

Variable Importance Rank	Entrez Gene ID	Gene Symbol	Gene Description	Gain	Cover	Frequency
1	3945	LDHB	lactate dehydrogenase B	0.70	0.22	0.06
2	51780	KDM3B	lysine (K)-specific demethylase 3B	0.13	0.12	0.07
3	29087	THYN1	thymocyte nuclear protein 1	0.04	0.02	0.06
4	23013	SPEN	spen homolog, transcriptional regulator (Drosophila)	0.04	0.02	0.04
5	9535	GMFG	glia maturation factor, gamma	0.01	0.04	0.06
6	55737	VPS35	vacuolar protein sorting 35 homolog (S. cerevisiae)	0.01	0.01	0.03
7	9559	VPS26A	vacuolar protein sorting 26 homolog A (S. pombe)	0.01	0.01	0.02
8	4694	NDUFA1	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 1, 7.5kDa	0.01	0.04	0.05
9	529	ATP6V1E1	ATPase, H+ transporting, lysosomal 31kDa, V1 subunit E1	0.01	0.06	0.03
10	9736	USP34	ubiquitin specific peptidase 34	0.01	0.01	0.04
11	317	APAF1	apoptotic peptidase activating factor 1	0.01	0.00	0.02
12	6625	SNRNP70	small nuclear ribonucleoprotein 70kDa (U1)	0.01	0.05	0.02
13	3054	HCFC1	host cell factor C1 (VP16-accessory protein)	0.00	0.01	0.05
14	997	CDC34	cell division cycle 34 homolog (S. cerevisiae)	0.00	0.00	0.02
15	3108	HLA-DMA	major histocompatibility complex, class II, DM alpha	0.00	0.06	0.05
16	3028	HSD17B10	hydroxysteroid (17-beta) dehydrogenase 10	0.00	0.02	0.03
17	23526	HMHA1	histocompatibility (minor) HA-1	0.00	0.04	0.02
18	10695	CNPY3	canopy 3 homolog (zebrafish)	0.00	0.04	0.02
19	51283	BFAR	bifunctional apoptosis regulator	0.00	0.04	0.06
20	27095	TRAPPC3	trafficking protein particle complex 3	0.00	0.02	0.04
21	6166	RPL36AL	ribosomal protein L36a-like	0.00	0.02	0.04
22	7832	BTG2	BTG family, member 2	0.00	0.06	0.04
23	2976	GTF3C2	general transcription factor IIIC, polypeptide 2, beta 110kDa	0.00	0.00	0.03
24	5436	POLR2G	polymerase (RNA) II (DNA directed) polypeptide G	0.00	0.06	0.01
25	93487	MAPK1IP1L	mitogen-activated protein kinase 1 interacting protein 1-like	0.00	0.01	0.03
26	6742	SSBP1	single-stranded DNA binding protein 1, mitochondrial	0.00	0.02	0.04
27	16	AARS	alanyl-tRNA synthetase	0.00	0.00	0.02
28	29	ABR	active BCR-related	0.00	0.00	0.02

The "AD vs mixed control" classification model uses 28 predictive features for patient diagnosis. Variable importance was calculated and provided above to highlight the most predictive features.

Supplementary Table 5.2: GSEA results on the "AD vs mixed control" predictive features

Pathway	Database Source	Gene Overlap	p-value	q-value
WNT ligand biogenesis and trafficking	Reactome	VPS35; VPS26A	9.56E-04	3.82E-02
Alzheimer disease - Homo sapiens (human)	KEGG	HSD17B10; APAF1; NDUFA1	3.70E-03	5.19E-02
Herpes simplex infection - Homo sapiens (human)	KEGG	CDC34; HCFC1; HLA-DMA	4.61E-03	5.19E-02
Huntington disease - Homo sapiens (human)	KEGG	POLR2G; APAF1; NDUFA1	5.19E-03	5.19E-02

GSEA was performed on the 28 predictive genes of the "AD vs mixed control" classification model. The most significant results ($p \leq 0.05$) are provided above.